THE EMERGENCE OF SLIPPAGE-TYPE EDITING IN AN EVOLUTION EXPERIMENT

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in Cellular and Molecular Biology

at the

University of Canterbury

by

Alicia Sook Wei Lai

UNIVERSITY OF CANTERBURY

2016

Table of contents

Acknowledgements

I wish to thank my supervisor Ant for:

- Accepting me into his lab group where I have met some of the most amazing people
- Sending me around the globe to attend some of the most amazing conferences and meetings
- Putting up with my silly antics and constant 'stalking'
- Constantly telling me that I did good but I can always do better
- Relentlessly questioning me on where I see myself in 5 years' times

In all seriousness, I thank you from the bottom of my heart for the endless help, guidance, support and knowledge you have provided me throughout my PhD. It has been a crazy ride, but the journey doesn't stop here, there're still a lot of stones left unturned!

I would also like to acknowledge everyone in the molecular biology lab group especially Ryan, aka the brains (not forgetting beauty)! Ryan, I am forever grateful for your guidance and knowledge, your witty remarks and your everlasting friendship. Your intelligence never ceases to amaze me!

Alannah, you have always been there for me, through thick and thin even though I've been bossing you around the lab and telling you off for being…you (if you're reading this right now, you know that I tend to over exaggerate). In all honesty, you've been one of my biggest supporters and you've always had my back! Thank you so much for putting up with through these stressful times, you're da bomb!

Nicole, I'm grateful for your fruitful input and discussions on the delta-bitscore analysis. I sincerely hope that one day you'll receive an email from me that doesn't start with something like "could you do me a favour".

I would also like to thank Nellie for bossing me around. Even though you started out as one my 'adopted' summer students, your constant questioning of my intelligence and leadership encouraged me to work harder to please you. Our endless bickering over scientific issues helped me broaden my thinking and I sincerely thank you for that.

Also, big thanks to Jack, Ren and not forgetting Paul for the great discussions and advice over the years. Not forgetting Duncan who worked with me on my project over the summer! You've been a great help and friend. To everyone else in the lab and those who helped me over the past years, thank you – your support made this thesis possible.

To my dearest Mum and Dad, thank you for everything that you've done for me. Without your love, help and support, I would not have made it to New Zealand to pursue my PhD and life goals. I apologise for spending all my time on my thesis and not returning home to Malaysia to visit. I promise to come home soon to spend time with you and share my experiences with you folks. I am proud to say that I am who I am now, and I am where I am, because of you.

"If I have seen further than others, it is by standing upon the shoulders of giants."

- Isaac Newton

Abstract

RNA editing, the correction of genomic errors by altering the sequence of messenger RNA, has evolved multiple times independently, but it remains unclear exactly how such a complex yet low-fidelity process has evolved. A model for the emergence of RNA editing proposed that RNA editing activity pre-exists, but there is no substrate for editing activity to act upon. Subsequently, mutation creates editable nucleotide sites, which may be fixed by genetic drift making RNA editing indispensable for expression of functional genes. We sought to test this model by asking whether editing-type processes can evolve under experimental conditions designed to maximize the effect of genetic drift.

Previous work in our group has shown that, in the bacterial endosymbiont *Buchnera*, RNA polymerase slips at long poly(A/T) tracts, leading to stochastic incorporation or removal of As or Us in the nascent messenger RNA. This results in a heterogeneous population of mRNAs. Slippage-type editing was shown to correct errors in genes which had acquired natural frameshift mutations, but this may in turn reduce expression efficiency of genes with intact reading frames. It would appear that these editing processes in *Buchnera* may not always be beneficial thus raising a question: if editing-type processes do not offer any inherent advantage, how did this process emerge and why has it persisted after millions of years?

In a mutation accumulation (MA) experiment, we subjected *Escherichia coli* populations to daily single-cell bottlenecks, mimicking the genetic background of *Buchnera*. After approximately 4,000 generations, a general loss of fitness was observed while one of the lineages succumbed to mutational meltdown. Genome sequencing revealed the accumulation of mutations and the emergence of 22 frameshift mutations that require slippage-type editing

for the production of functional proteins. We then introduced one of the frameshift mutations into wild-type *E. coli* and demonstrated that RNA polymerase slippage results in a loss of fitness. Further to this, we also conducted delta-bitscore (DBS) analyses to identify functional perturbation following mutation accumulation and we showed loss of fitness may be attributed to the loss of protein function in the bottlenecked lineage that succumbed to mutational meltdown.

We subsequently assessed the impact of RNA polymerase slippage on gene expression and protein production utilising GFP reporter systems. We present data showing that slippage-type editing rescues frameshift mutations but that protein production is reduced. Our results support the hypothesis that, under conditions favouring genetic drift, editing-type processes may readily emerge. To our knowledge, this is the first experimental demonstration of the evolutionary drivers for the emergence of RNA editing.

CHAPTER 1
Introduction

Rationale of research

Following the origin of life, the notion of evolution from simple to complex has been widely accepted as complexity is generally equated with biological improvement or advancement (Lynch, 2007a). However, complex low-fidelity and multi-layered processes such as RNA editing may equally emerge under conditions favouring significant genetic drift and inefficient selection (Gray et al., 2010; Lynch, 2007a; Stoltzfus, 1999). Slippage-based editing has been characterised in obligated insect endosymbionts and slippage was demonstrated to correct errors in genes which have acquired natural frameshift mutations, but may in turn reduce expression efficiency of genes with intact reading frames (Tamas et al., 2008). This is consistent with the 3-step drift model of RNA editing by Covello and Gray (1993) where complex editing-type processes may evolve despite offering no inherent advantage. Although RNA editing has been described and a role for genetic drift in the emergence of editing-type processes has been proposed (Covello and Gray, 1993), till date, there are no experimental demonstrations to support this hypothesis. We therefore sought to experimentally test the drift model for the origin of RNA editing and assess how this complex, yet inherently disadvantageous process affects cell fitness, gene expression and protein production.

In Chapter 1 of this thesis, I will be covering the following concepts: i) Molecular complexity: I will discuss whether complexity may evolve non-adaptively and address the

debatable view that evolution is only directed towards favourable and advantageous traits. For the purposes of this thesis, complexity is defined as an increasing number of steps interdependencies or componets as a part of a molecular systems, relative to a more straightforward way to perform the same task ii) RNA editing: a complex and superficially unnecessary process *per se*. I will review the concept of RNA editing, its suggested role as an error correction system, its source of transcriptomic and proteomic diversification, and how this complex process may have evolved under neutral conditions. iii) Muller's ratchet: this phenomenon plays a major role in driving mutational trajectories in small asexual populations by allowing the accumulation of slightly deleterious mutations. I will discuss the relevance of this Muller's ratchet for my mutation accumulation experiment. The neutral model for the evolution of RNA editing by Covello and Gray (1993) is fundamental to this study, and I will begin with a discussion on this model.

Three-step neutral model for the evolution of RNA editing

RNA editing is a process by which a RNA sequence is altered with respect to the corresponding DNA or RNA that encodes it (Benne et al., 1986). This results in mRNA transcripts that do not reflect the original template. In many, but not all cases, the outcome of RNA editing is translatable transcripts that would not otherwise be functional in the absence of editing (Linton et al., 1992, 1997). It would seem logical to produce the correct, functional transcript initially, so why instead do we see a requirement for editing?

The neutral model for the evolution of RNA editing is an example of the cart-before-horse scenario where the capacity for editing (cart) arises before the emergence of the substrate (horse), that the enzyme acts upon. Stoltzfus (1999) suggested that the evolution of the editing machinery can be explained by tinkering. François Jacob (1977) suggested that a

'tinkerer' gives his materials unexpected functions to produce a new object as a result of a series of contingent events, and in addition, none of the materials have a definite function (Levi-Strauss 1962). In other words, evolution is a matter of tinkering where novel structures and more elaborate functions are produced from the alteration of existing compounds and not by scratch (Jacob, 1977). This description is consistent with the model proposed for the origins of RNA editing where RNA polymerase may perform genome error correction (in the case of RNA editing) besides its primary role in transcription.

Covello and Gray (1993) proposed a general three-step general model for the evolution of RNA editing systems (Box 1.1), and the model has been extended by Stoltzfus (1999) to include kinetoplastid editing. Under this model, RNA editing activity pre-exists but there is no substrate for the enzyme to act upon. Subsequently, random mutation creates editable nucleotide sites. Upon fixation of these mutations by genetic drift, RNA editing becomes an indispensable part of the genetic information pathway and is thus maintained by natural selection. This model of RNA editing evolution is an example of a more general scheme of constructive neutral evolution (CNE) that accounts for a series of steps giving rise to novel structures and mechanisms (Stoltzfus, 1999). In this chapter, I will further discuss this model of CNE and provide examples of complex systems which such as RNA editing systems that appear to have evolved in a neutral fashion.

Box 1.1 | 3-step model for the evolution of RNA editing

Step 1: The appearance of the RNA editing machinery before there is need for editing

Step 2: Mutation at editable nucleotide positions and fixation by drift

Step 3: Maintenance of RNA editing by selection

Evolution from simple to complex, or complex to simple?

The common perception of evolution posits that the process generally proceeds from simple to complex, towards increases in size (Bonner, 1988), complexity (Bonner, 1988; McShea, 1996) and diversity (Foote and Gould, 1992). In favour of evolution towards greater complexity, Lenski and colleagues demonstrated that complex functions were built upon simpler functions and structures using digital organisms, computer programs that self-replicate, mutate, compete and evolve (Lenski et al., 2003). Yet, there is no inherent reason that evolution should be directed towards complexification. Simplification may equally occur, as pointed out by Szathmáry and Smith (1995) and early organisms were more complex than previously thought. Extant organisms such as sponges, comb jellies, and placozoans may have evolved through gene loss, and perhaps became simplified from a more biological complex ancestor (Fortunato et al., 2014; Mendivil Ramos et al., 2012; O'Malley et al., 2016), providing evidence that evolution is not directed towards complexification alone. Additionally, many examples of the evolution of parasitic and symbiotic organisms suggest that prokaryotes are not simplified *per se* but streamlined (Lynch, 2006) by selection for small genomes (Andersson and Kurland, 1998; Ham et al., 2003; Mccutcheon and Moran, 2012; Moran, 2002; Oshima et al., 2004) and for fast replication (*r*-selection) (Blattner et al., 1997; Levinson and Gutman, 1987). Mitochondria, the ubiquitous organelle responsible for energy conversion in the majority of eukaryotes, and chloroplast, the organelle capable of photosynthesis, arose as bacterial endosymbionts (Gray, 2012a; Gray and Doolittle, 1982; Gray et al., 1999; Moreira et al., 2000), and are the ultimate realisation of reductive evolution.

Complexity seems to be a by-product of evolutionary processes (Lynch, 2007a; Poole et al., 2003; Smith and Szathmáry, 1995) and the increase in complexity may have been achieved as

a result of a series of major evolutionary transitions (Szathmáry and Smith, 1995). Natural selection, genetic drift, mutation, and gene flow are the basic mechanisms of evolution, and evolutionary changes are not dictated by a single cause but instead as a consequence of a specific combination of these mechanisms and are contingent on previous events (Lewontin, 2002). Thus, based on what is known about evolution to date, evolution may not be governed by a single evolutionary mechanism, and although adaptation of microbes by natural selection has been well documented (Barrick et al., 2009; Elena and Lenski, 2003; Lenski, 2017), we cannot rule out the importance of chance, mutation, and recombination in evolution. To help explain the rationale for this thesis, I will now elaborate on the theory of non-adaptive evolution, termed constructive neutral evolution.

Constructive neutral evolution

One of the key tenets of neo-Darwinism is that change (or in the case of this thesis, the increase in molecular complexity) evolves step-by-step through a gradual evolutionary process driven by selection (Zhang 2010). Gould (1997) however, argued that the evolution of complexity follows a "drunkard's walk" model. This "drunkard's walk" model is based on the idea of evolutionary contingency, where Gould claimed that the path towards complexity occurs by chance (Gould, 1997). This idea is conceivable, as that there are other evolutionary forces besides natural selection: genetic drift, mutation, and gene flow are also capable of driving or changing evolutionary trajectories. Gould and Lewontin (1979) pointed out that although most phenotypic traits that contribute towards fitness are undoubtedly selected for, it is not necessarily right to assume that all phenotypic traits are a product of selection. The authors claimed that some phenotypic traits that do not provide any advantage, termed "spandrels" may be fixed in spite of natural selection (Gould and Lewontin, 1979). The question is apart from natural selection, can molecular complexity evolve by chance?

The concept of constructive neutral evolution (CNE) was proposed by Arlin Stoltzfus (1999) as an extension to the neutral model of the evolution of RNA editing proposed by Covello and Gray (1993). Stoltzfus stated that neutral changes in succession with no fitness attribution account for a complex series of steps giving rise to novel structure, functions and operations. These assertions of the neutral evolution of complex processes (Covello and Gray 1993; Stoltzfus 1999) are in line with François Jacob's (1977) hypothesis of evolution by means of tinkering, where complexity emerges by historical contingency without blueprints. Gray and colleagues (2010) added that fortuitous molecular interaction between molecules can lead to a fixed dependency if this interaction compensates for a mutation that was otherwise lethal, in a process termed pre-suppression (Figure 1.1). In addition, small finite populations are subjected to fixation of neutral or slightly deleterious mutations resulting in an inevitable increase in complexity (Fernández and Lynch, 2011; Lynch, 2007a, 2007b). Therefore, an increase in dependency is unavoidable if there are more ways for dependence to increase than decrease, ultimately leading to irremediable complexity. This process of pre-suppression allows unnecessary dependencies to accumulate, resulting in a large bureaucratic system that essentially does a single task through multiple processes (Gray et al., 2010). Thus, CNE may provide an alternative explanation of evolutionary changes other than selection and this idea of CNE may be evoked to explain the evolution of fortuitous complexity (Lukeš et al., 2011; Stoltzfus, 2012). To elaborate on how CNE is relevant to this thesis, I will now discuss how CNE may explain the non-adaptive emergence of complex molecular systems, as reasoned in the literature.

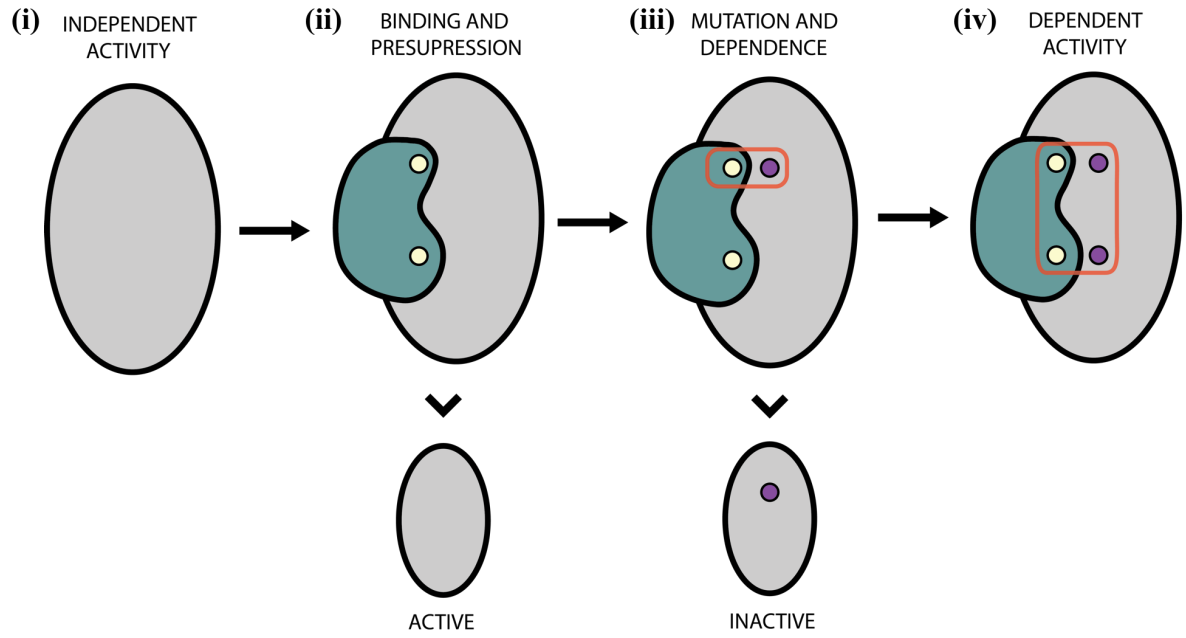

Figure 1.1 | The neutral evolution of complexity through the action of pre-suppression. Schematic depicts: **(i)** Two generic components (A and B) which are initially functionally independent of each other. **(ii)** A fortuitous, yet neutral, interaction between A and B presuppresses mutations in A. At this point, A is still functional independent of B. **(iii)** A mutation in A then occurs, rendering A functionally inactive, but due to pre-suppression by B, A is active but now dependent on the presence of B. **(iv)** Subsequent mutations render A further functionally dependent on B due to continued pre-supression of mutations. This complexity (where A is reliant on B for functionality) is increased by further mutations. This provides an example of constructive neutral evolution acting as a directional force that drives increasing complexity without positive selection. Figure modified from Gray et al., (2010).

Molecular complexity may arise through non-adaptive processes

There have been a number of cases where molecular complexity has been proposed to have arisen through non-adaptive processes, such as codon reassignment and splicing. Osawa and Jukes (1989) originally proposed codon reassignment as a process where, during evolution, loss of a codon and subsequent loss of the corresponding tRNA carrying the anticodon may

occur. At a later period of evolution, the lost codon reappears through mutation, along with the tRNA that translates it, also as a result of mutation. The tRNA now has the capacity to translate both the reappeared codon and the lost codon, therefore reassigning or capturing the codon. This codon reassignment follows the codon capture rule, where a series of non-disruptive neutral changes occur in the codon, ultimately changing the genetic code (Osawa and Jukes, 1989; Osawa et al., 1992). Stolzfus (1999) extended this idea and reasoned that this codon capture theory follows neutral evolution, whereby the "temporary" disappearance of an amino acid codon may arise by mutation of every occurrence of that codon to an alternative codon, combined with subsequent loss of its corresponding tRNA. This results in a deleterious allele due to this codon no longer coding for its original amino acid, until the reappearance of an appropriate tRNA that could translate the lost the codon. Spontaneous duplications of tRNA genes and mutations that change tRNA specificity could then generate a new tRNA for this lost codon during the course of evolution, which may result in the codon coding for either the same or a different amino acid (Osawa et al., 1992). These conditions allow the lost codon and the potentially altered genetic code to accumulate and be fixed by drift. The evolutionary origins of the reassigned codon are therefore more complex than the original codon, with an additional codon loss and capture step. In a more recent study, Steven Massey (2015) extended the hypothesis of the non-adaptive evolution of the genetic code and proposed that superior genetic codes with higher error minimization compared to the standard genetic code (SGC) can emerge neutrally via genetic code expansion. The emergence of these beneficial traits termed "pseudaptations" suggests that not all traits that result in an increase of fitness have arisen by the direct action of natural selection (Massey, 2015).

Further to this, CNE has been suggested as an explanation for the increase of complexity in splicing. Splicing is the editing of precursor mRNA transcripts that is catalysed by the

spliceosome, a complex multimegadalton ribonucleoprotein (RNP) machine assembled from five small nuclear RNAs (snRNAs): U1, U2, U3, U4, U5 and U6, and hundreds of proteins (Will and Lührmann, 2011). It is generally agreed that eukaryotic spliceosomal snRNAs are derived from group II introns, a class of self-catalytic ribozymes that catalyse their own removal with a similar splicing reaction to the spliceosome. Extensive similarities are seen between the self-splicing group II introns and the splicesomal snRNAs in their chemical mechanism of cleavage, structurally analogous regions, transesterification steps (Cech, 1986; Copertino and Hallick, 1993; Sharp, 1985) and their phylogenetic distribution (Cavalier-Smith, 1991), suggesting a common origin between. Domain-swapping experiments of the ID3 (inhibitor of DNA binding 3) subdomain of a group II intron indicates that a proportion of structural elements of group II introns and snRNAs are functionally analogous (Hetzer et al., 1997; Shukla and Padgett, 2002). It was suggested that a proportion of group II introns fragmented early on, yielding primordial snRNAs that could facilitate splicing in *trans*, which then allowed the degradation of other introns (Copertino and Hallick, 1993). The fragmentation of these introns would be ratchet-like because correct reassembly of the fragmented introns would be rare (Lukeš et al., 2011; Stoltzfus, 1999). Pre-existing RNA-binding proteins such as the spliceosomal factors may interact by chance and the pre-suppression action of the RNA-protein binding therefore increases the complexity of the extant spliceosome by an unselected step. In addition, if early eukaryotes had small populations, then fixation of neutral or slightly deleterious mutations may have contributed to the complexity of the extant multicomponent splicesosome (Lynch, 2007a, 2007b).

Finnigan and colleagues (2012) provided experimental evidence as to how loss of function rather than a gain in function drives the emergence of cellular complexity in a non-adaptive manner. The vacuolar H^+-ATPase (V-ATPase) is the universal rotary proton pump of

eukaryotes essential for vesicular trafficking (Forgac, 2007) and the *Saccharomyces cerevisiae* V-ATPase ring consists of a Vma16 subunit, 4 Vma3 subunits and one Vma11 subunit, where the Vma3 and Vma11 subunits are paralogs and were products of a duplication of an ancestral gene, Anc.3-11. If these subunits are products of selection, we would predict that there would be an improved performance or additional function compared to its ancestral state. However, it was shown that these paralogous subunits lost the ability to bind to the both interfaces of Vma16, yet through ancestral gene reconstruction, the ancestral derived subunits were capable of binding to both the interfaces of Vma16. These results suggested that during the course of evolution, the increased complexity of the hexameric ring in the extant yeast did not contribute towards the gain of new functions and the duplicated subunits may have been fixed by chance. Thus, the accumulation of more subunits and patterns of the assembly of the V_0 protein hexamaric ring does not confer improved or novel functions on the protein complex, suggesting that complexity may arise in a neutral manner.

To summarise, these studies suggest that, as mutations that do not cause any phenotypic changes emerge, complexity may arise as a side effect in the absence of selection. This process of non-adaptive evolution of complexity, termed CNE (Stoltzfus, 1999), is a ratchet-like, unidirectional process that can increase complexity without providing any additional advantage. Once this unidirectional ratchet is established, selection may then play a role in preventing this directionality from being reversible, making complexity irremediable (Gray et al., 2010). Once again building on the hypothesis on evolution and tinkering, evolution does not only produce novelties from scratch but novelties may also be generated through alterations of existing elements producing objects of increasing complexity in the absence of selection (Jacob, 1977) as observed in CNE.

RNA editing as an example of non-adaptive evolution

RNA processing systems such as editing have been suggested to be exemplary of superfluous complexity (Gray et al., 2010). RNA editing is distinct from the events of intron excision, RNA splicing, 5'-capping, 3'-polyadenylation, pseudouridine formation, methylation, and degradation (Maniatis and Reed, 2002). RNA editing covers a series of biochemical and enzymatic events that result in nucleotide alterations of a RNA sequence relative to its corresponding DNA template (Gray, 2012b). Benne and colleagues first introduced the term RNA editing more than 30 years ago when they discovered sequence alteration of RNA molecules in the mitochondria of trypanosomes (Benne et al., 1986). This phenomenon of the insertion and deletion of uridine (U) is a unique post-transcriptional RNA modification process that involves the addition and removal of uridine residues at precise sites, usually within coding regions, generating functional proteins (Aphasizhev et al., 2002; Decker and Sollner-Webb, 1990; Simpson et al., 2003). Studies have shown that editing of mRNAs corrects genomically encoded frameshifts, creates initiation and termination codons for mitochondrial translation, and in extensively edited mRNAs, the formation of complete reading frames from nonsense sequences (Estévez and Simpson, 1999; Hajduk et al., 1993). The sequence alteration of RNA molecules has been found in the tRNA, rRNA, mRNA, and miRNA molecules of eukaryotes (Brennicke et al., 1999) and more recently, tRNA molecules of archaea (Randau et al., 2009) (Table 1.1).

RNA editing is a non-heritable change to the RNA which has been proposed to result in: 1) An increase in genetic variation (Landweber and Gilbert, 1993), and transcriptional repertoire and proteome diversity (Pullirsch and Jantsch, 2010) under certain circumstances; 2) Additional regulation of gene expression (Stuart et al., 1997) and; 3) Fixation of mutations in the mitochondrial genome of kinetoplastan protozoa under anaerobic environments (Cavalier-

Smith, 1997). Although RNA editing has been proposed to have an apparent function, as stated above, there is insufficient evidence that favour any of the proposed functions. Alternatively, RNA editing has been suggested to have simply evolved to either correct an error at the genome level producing translatable mRNA transcripts (Shaw et al., 1988; Simpson et al., 2000, 2003), or modify a functional mRNA to generate alternative proteins (Nishikura, 2010).

RNA editing was first discovered in the kinoplast mitochondrion of trypanosome, a unicellular kinetoplastid protozoa, where editing occurs in approximately 12 out of 18 mRNA transcripts (Estévez and Simpson, 1999). RNA editing through the insertion or deletion of uracil (U) residues in mitochondrial mRNA is mediated by guide RNAs (gRNAs), and involves endonuclease cleavage of the editing site mediated by these gRNAs, followed by the addition, or occasionally deletion of U's from the 3' end of the 5' fragment and re-ligation (Blum et al., 1990). This phenomenon of U-insertion or deletion during gene expression does not destroy the genetic message as expected, instead, the edited transcripts encode functional proteins whereby the insertion/deletion of U's corrects frameshifts (Aphasizhev and Aphasizheva 2011). The evolutionary origins of U-insertion or deletion editing have been highly debated, as some propose that editing is a relic of the RNA world (Gilbert, 1986), having been involved in primordial error correction. However, the narrow phylogenetic distribution (Deschamps et al., 2011), where this form of editing is limited to the kinetoplastid protist lineage, suggests that RNA editing is unlikely to be an artefact of the RNA world but instead appears to be derived traits within the lineages in which they are found (Gray, 2003, 2012b).

Table 1.1 | Different types of RNA editing

Editing type	Organisms	Affecting	Genetic system	Reference
U insertion/deletion	Kinetoplastid protozoa	mRNAs	Mitochondrial	Benne et al., 1986
C-to-U conversion	Mammals Plants	mRNAs mRNAs	Nuclear Mitochondrial and chloroplast	Powell et al., 1987 Covello and Gray, 1989; Gray and Covello, 1993; Hiesel et al., 1989
U-to-C conversion	Land plants and placozoa Archaea	mRNAs tRNAs	Mitochondrial and chloroplast Bacterial	Burger et al., 2009; Kugita et al., 2003 Randau et al., 2009
A-to-I conversion	Metazoa	mRNAs, tRNAs, miRNAs, and viral RNAs	Nuclear	Nishikura, 2010
N replacement	*Acanthamoeba* and chytridiomycete fungi	tRNA 5' acceptor stem	Mitochondrial	Laforest et al., 1997; Lonergan and Gray, 1993
N insertion	Slime molds	mRNAs, tRNAs and SSU rRNAs	Mitochondrial	Visomirski-Robic and Gott, 1995
Multiple substitutions	Dinoflagellate	mRNAs and rRNAs	Mitochondrial and chloroplast	Lin et al., 2002; Zauner et al., 2004

Logically, in trans-acting gRNA-mediated uridine insertion or deletion editing (Simpson et al. 2000), the U-insertion or deletion sites should arise prior to the gRNAs that correct them. However, because the first step in U insertion or deletion editing requires transcript cleavage mediated by gRNAs, gRNAs must have appeared before the errors they correct were fixed in the population. In other words, duplication and anti-sense transcription of a gene segment that gives rise to a gRNA gene must precede the deletion of T, because duplication of the gene after to the T deletion would give rise to a gRNA that lacks this nucleotide position (Stoltzfus, 1999). Because of this, a T deletion at the DNA level could be tolerated, allowing this mutation to persist and spread by genetic drift (Stoltzfus, 1999). A constructive neutral evolution (CNE) origin of U-insertion/deletion RNA editing proposes that the enzymatic machinery (RNA helicase, endoribonuclease, terminal uridylyl transferase and RNA ligase capable of U-insertion or deletions) normally present in mitochondria as components of other RNA-processing pathways must have pre-dated the first edited site. These enzymes may have emerged by duplication and divergence of genes with other functions, and by chance, interacted with the gRNA precursors allowing mutations in the genome to be corrected at the RNA level. RNA editing was proposed to play a regulatory role in kinetoplastids (Stuart, 1991) thus providing a selective advantage. However, this does not necessarily mean that RNA editing evolved to function as a gene expression regulator, but rather that once RNA editing was established it could have been repurposed to serve a regulatory role (Covello and Gray, 1993).

The initial discovery of RNA editing in protists reported U-to-A, U-to-G and A-to-G conversions in the first three nucleotides of the 5′ half acceptor stems of mitochondrial tRNAs which restored canonical base pairing (Lonergan and Gray, 1993). This nucleotide exchange process was also found in *Splizellomyces punctatus* (Laforest et al., 1997) and

Hyaloraphidium curvatum (Forget et al., 2002). A total of 23 RNA editing events of this type were reported to reconstitute proper base pairing in 13 out of a total of 16 tRNAs encoded in the mtDNA of *Acanthamoeba castellanii* (Lonergan and Gray, 1993; Price and Gray, 1999). This however, raises questions as to why most mtDNA-encoded tRNAs require editing even though the process of nucleotide conversions is not as simple as a biochemical transformation of the bases. Under the CNE model, an enzyme capable of removing the first 3 nucleotides, and subsequently incorporating the missing nucleotides emerge prior to the mutation fixation. Although the putative nuclease component of the editing activity has yet to be characterised, the polymerase component has been identified as a Thg1-like protein (TLP) and its activity has been demonstrated in *A. castellanii* (Price and Gray, 1999) and *Spizellomyces punctatus* (Bullerwell and Gray, 2005). Genes with similar sequences to Thg1 have been identified in all three domains of life (Abad et al., 2011; Heinemann et al., 2010; Rao et al., 2011), suggesting that mitochondrial tRNA-editing activity could readily occur by duplication and divergence of TLP genes (Gray, 2013). The emergence of mitochondrial TLP relaxes the evolutionary constraints on the editable first 3 nucleotides of the mtDNA-encoded tRNA genes enabling mutations to accumulate and be fixed by chance at these sites in the presence of editing. Once the interaction between the tRNA-editing system and editable sites has been established, the mismatch mutations may spread to other mitochondrial tRNA genes (those that could be corrected by editing) and consequently selection then acts to prevent the loss of the editing system.

To summarise, the concept of constructive neutral evolution (CNE) was originally proposed to provide an alternative hypothesis to the adaptive evolution of complex systems such as RNA editing, gene-scrambling and splicing (Stoltzfus, 1999). Systems that evolved non-adaptively may be subsequently co-opted for secondary functions under positive selection

(Covello and Gray, 1993), and a further example of such a system is seen in slippage-type editing. I will now elaborate on slippage-type editing, and how the evolution of this process can be explained by CNE.

Slippage-type editing is a complex process analogous to RNA editing

I have discussed how complex systems such as RNA editing may have evolved non-adaptively through historical contingency. RNA editing has been historically and broadly defined as a post-transcriptional process that modifies RNA molecules relative to its encoding DNA template (as discussed earlier), and although slippage-type editing is not a post-transcriptional process, the final outcome is analogous to that of RNA editing, where changes are made at the RNA level, ultimately resulting in RNA transcripts that do not reflect its coding gene. Slippage-type editing generally changes the genetic code at the RNA level, affecting gene expression and its origin could reveal some fundamental and generalisable insights into the origins of complex editing-type processes.

Accurate transmission of genetic information during transcription requires that RNA polymerase maintains the correct register of the DNA-RNA hybrid. A shift in the register of the RNA with the DNA template would result in misincorporation of nucleotides. In cases where the DNA-RNA hybrid consists of a homopolymeric tract of As or T/Us (AT/AU hydrogen bond is thermodynamically weaker compared to GC) (discussed below), RNA polymerase slippage and subsequent elongation without translocation of the active site was observed (Figure 1.2) (Anikin et al., 2010). The stuttering of RNA polymerase where RNA polymerase transcribes a nucleotide several times in the absence of translocation (movement of the DNA template and DNA-RNA hybrid through RNA polymerase) and transcript slippage results in the incorporation, or occasionally removal, of extra nucleotides into the

nascent RNA transcript that were not encoded in the DNA template (Chamberlin and Berg, 1962). Slippage-type editing, also known as transcriptional slippage has been observed during all phases of the transcription cycle, including initiation, elongation, and termination (Anikin et al., 2010).

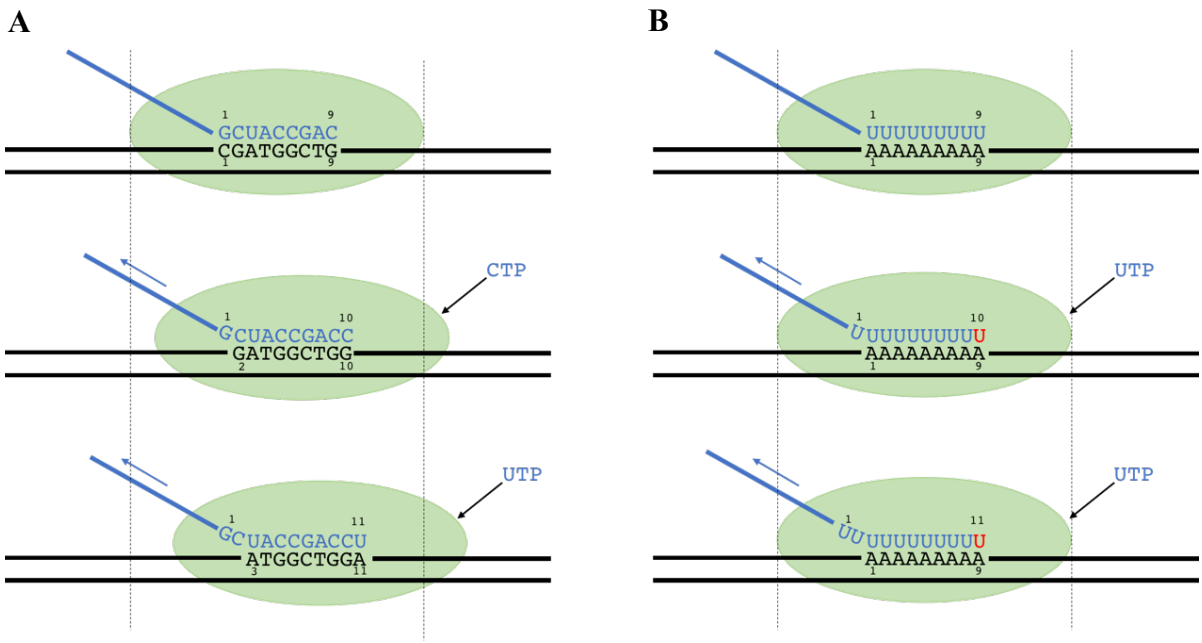

Figure 1.2 | During regular transcription, RNA polymerase faithfully transcribes the DNA template while long homopolymeric tracts promote RNA polymerase slippage. **A)** During normal transcript elongation, faithful nucleotide incorporation is followed by RNA polymerase movement along the DNA template. The RNA polymerase complex maintains the correct register of the DNA-RNA hybrid, enabling high fidelity transcription. Under these conditions, polymerisation is coupled to translocation. **B)** However, during transcription when RNA polymerase encounters a long tract of homopolymeric As, RNA polymerase stutters, and the transcript slips along the template thus introducing or removing non-templated adenine residue(s) to the transcript.

The phenomenon of transcriptional slippage was first proposed by Chamberlin and Berg when they observed the event of RNA polymerase 'slipping' during transcription (Chamberlin and Berg, 1962). *In vitro* transcription of denatured calf thymus DNA by *E. coli* RNA polymerase in the presence of ATP as the sole substrate yielded a transcript of polyadenylic acid (Chamberlin and Berg, 1962). The length of the transcript was longer than the template and the addition of any of the other three NTP substrates resulted in the inhibition of polyadenylic acid production. The authors suggested that RNA polymerase reiteratively transcribes poly(dT) tracts by repeated cycles of melting the DNA-RNA hybrid, slippage of the transcript, rehybridisation and subsequent incorporation of AMP (Chamberlin and Berg, 1962). Transcriptional slippage is therefore defined as the process of RNA polymerase slipping on long poly(A) or poly(T) tracts, resulting in the stochastic incorporation or removal of one or more nucleotides to the nascent RNA. This generates a heterogeneous pool of mRNAs with varying lengths and reading frames (Wagner et al., 1990, Tamas et al., 2008).

The ability of RNA polymerase to slip on homopolymeric sequences above a certain length relies on the length and stability of the DNA template and nascent RNA duplex within the polymerase (Parks et al., 2014; Zhou et al., 2013). Studies on the DNA-RNA hybrid in maintaining the correct registry of the ternary elongation complex (TEC, composed of RNA polymerase, DNA template and RNA transcript) showed that the DNA-RNA hybrid is 8-9 bp in length in *E. coli* (Nudler et al., 1997). Wagner and colleagues (1990) suggested that homopolymeric A or T tracts of 8 bases or longer are prone to RNAP slippage. The authors proposed that because A-U bonds are weaker than their G-C counterparts, the A-U hybrid may be melted at physiological conditions leading to dissociation or slippage of the nascent RNA (Wagner et al., 1990). This idea of a thermodynamically weak A-U duplex was

supported by studies on transcription elongation properties in vaccinia virus, where it was reported that the DNA-RNA interactions that stabilise the nascent RNA were weaker when polyU occupies most of the nascent RNA binding pocket of the RNA polymerase (Deng and Shuman, 1997). Overall, RNA polymerase slippage has been shown to occur more frequently on longer tracts of polyA/T (Penno et al., 2015; Uptain et al., 1997; Wagner et al., 1990).

Studies on genomes of free-living *E. coli* have shown that slippage-prone homopolymeric tracts are under-represented (Baranov et al., 2005). Transcription of these slippage-prone tracts is expected to result in aberrant and non-functional proteins that could be detrimental to the cells, and therefore these tracts are highly selected against under conditions of strong selection (Baranov et al., 2005). However, some studies on these slippage-prone homopolymeric tracts in prokaryotes have shown that these tracts may play a role as gene expression regulators (Orsi et al., 2010; Turnbough, 2011). I will now discuss the role of homopolymeric tracts in regulating gene expression in prokaryotes.

Slippage-type editing plays a role in gene expression regulation

RNA polymerase slippage is likely to produce aberrant products, yet slippage-type editing has been shown to have regulatory roles in *E. coli pyrBI* and *codBA* operons (Liu et al., 1994; Qi and Turnbough Jr, 1995). In *E. coli*, the *pyrBI* operon plays a part in the *de novo* synthesis of pyrimidine nucleotides. This operon consists of the *pyrB* and *pyrI* structural genes which encode the catalytic and regulatory subunits of the pyrimidine biosynthetic allosteric enzyme, aspartate transcarbamylase (EC 2.1.3.2). The expression of the *pyrBI* operon is mediated by cellular levels of the pyrimidine pool through UTP-sensitive slippage-type editing within the initially transcribed region: AATTTG of the pyrBI promoter (Figure 1.3, position is indicated by an asterisk). RNA polymerase slippage is directed by three T-A base pairs in the initially

transcribed region of the pyrBI promoter which result in nascent transcripts with the $AAUUUU_n$ (where n=l to >30) sequence. The weak base-paring between AAA and UUU allows slippage between the DNA-RNA hybrid resulting in nucleotide shifts in the nascent strand. Under conditions where UTP levels are high, RNAP slippage adds non-templated uridine residues that prevent downstream extension of the nascent transcript to include structural genes, resulting in the reduction of *pyrBI* operon expression. However, when UTP levels are low, correct repositioning can occur without uridine addition allowing the addition of a guanine residue. The addition of guanine results in a more stable base-paring preventing further RNAP slippage. These results suggest that regulation of gene expression of *pyrBI* by slippage-type editing functions as a response reductions in pyrimidine availability (Liu et al., 1994). The regulation of the *pyrBI* operon, therefore, prevents the unnecessary synthesis of aspartate transcarbamylase.

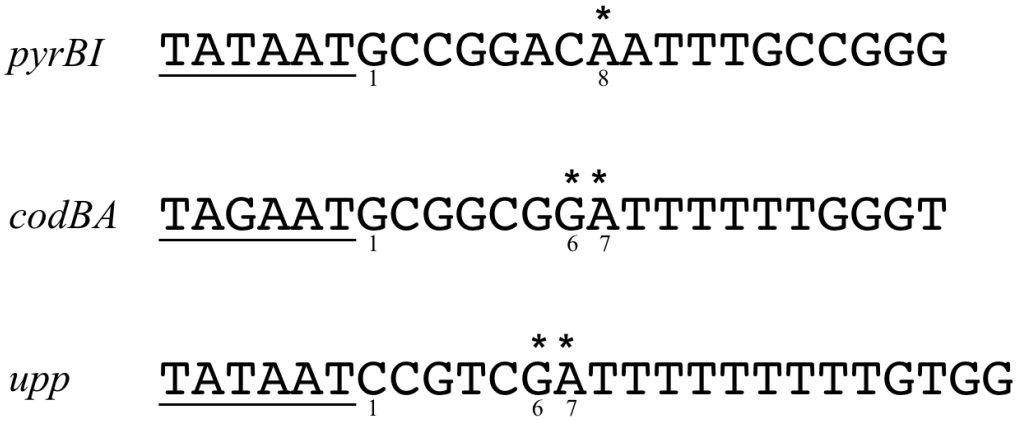

Figure 1.3 | Initially transcribed regions of the *pyrBI*, *codBA,* and *upp* promoters. Slippage-type editing plays a role in gene expression regulation relative to the cellular levels of pyrimidine. The –10 regions of the promoters are underlined and the major transcriptional start sites are indicated by asterisks. Start sites are identified according to their position downstream from the -10 region. The major start site of *pyrBI* is designated A8 while *codBA* and *upp* start sites are G6 and A7. Image adapted from Tu and Turnbough (1997).

Slippage-type editing has also been reported to play a role in the regulation of cytosine uptake and metabolism, where the expression of the *codBA* operon (Qi and Turnbough Jr, 1995) and *upp* (Tu and Turnbough, 1997) are regulated by the levels of UTP. Although the *pyrBI*, *codBA* and *upp* operons are regulated by the promoter region by the levels of UTP, the UTP levels control the selection of alternative transcriptional start sites of *codBA* and *upp* expression. The *codBA* operon consists of *codB* and *codA* which encode the enzymes cytosine permease and cytosine deaminase, respectively. In the presence of high UTP levels, the A8 transcriptional start site is favoured, resulting in slippage-type editing, and ultimately preventing high expression of the *codBA* structural genes. Low levels of UTP, on the other hand, inhibit the transcription initiation on the A8 position, favouring the G7 position instead (Figure 1.3, positions are indicated by asterisks). G7 does not engage in slippage-type editing and therefore permits the synthesis of full-length CodB and CodA. As a result, cytosine permease and cytosine deaminase are produced at high levels under the conditions of low UTP availability.

On the other hand, the *upp* gene encodes uracil phosphoribosyltransferase, a pyrimidine salvage enzyme that catalyses the formation of UMP from uracil and phosphoribosylpyrophosphate. The mechanism of *upp* expression regulation is similar to the *codBA* operon, where low intracellular levels of UTP favours transcription initiation at position G6 and high levels of UTP favour initiation at position A7, resulting in slippage. RNA polymerase slippage on the promoter region prevents transcript extension to the downstream *upp* sequences (Tu and Turnbough, 1997). As a whole, slippage-type editing has been suggested to be modulated to achieve physiologically relevant regulation of gene expression under certain conditions (Liu et al., 1994; Qi and Turnbough Jr, 1995).

Generation of multiple products from a single gene as a result of RNA polymerase slippage

While RNA polymerase slippage has been suggested to be disadvantageous for genes encoding a single functional protein product, slippage-type editing may be utilised for the synthesis of multiple products from a single gene (Baranov et al., 2005). The generation of multiple products from a single gene was proposed to provide variability and optimise the use of the limited genetic information in response to changes in environmental and growth conditions (Anikin et al., 2010).

The RNA polymerases of both *E. coli* and *Thermus thermophilus* have been shown to be slippage-prone, and their *dnaX* genes encode both the tau and gamma subunits of DNA polymerase III (Larsen et al., 2000). However, each bacterial species utilise a very different mechanism for generating the aforementioned subunits. The *dnaX* gene of *E. coli* consists of the A_6G motif sequence prone to ribosomal frameshifting (Box 1.2) (Flower and McHenry, 1990), while *T. thermophilus* consists of the A_9 motif which is prone to RNA polymerase slippage. Transcription of the slippage-prone tract generates a heterogeneous pool of mRNA in *T. thermophilus,* suggesting that ribosomal frameshifting was not being utilised for the production of the DNA polymerase III subunits. During transcription, *T. thermophilus* RNA polymerase slippage on the *dnaX* gene stochastically adds or removes A residues and when the number of As is equal to 9 or 9 + 3n, the full-length DNA polymerase III tau subunit is synthesised. However, when the number of As is anything other than 9 or 9 + 3n, the reading frame will be shifted in a manner where a stop codon will be generated downstream of the homopolymeric tract. The premature termination thus generates the gamma subunit of the DNA polymerase III subunit (Larsen et al., 2000).

Box 1.2 | Ribosomal frameshifting

Programmed ribosomal frameshifting (PRF) is a process whereby the ribosome shifts one nucleotide backwards (-1) or forwards (+1) into an overlapping reading frame and continues by translating a new amino acid sequence (Dinman, 2012). Although this process was historically associated with viruses (Jacks and Varmus, 1985; Jacks et al., 1988) and retrotransposons (Belcourt and Farabaugh, 1990), it is becoming increasingly apparent that PRF is widespread and has been documented in all kingdoms of life (Cobucci-Ponzano et al., 2012; Namy et al., 2004; Sharma et al., 2011).

In general, PRF is directed by cis-acting elements within the mRNAs that include a slippery shift site composed of a heptameric frameshift motif, a short spacer sequence, and a 3′ stimulatory element, typically a mRNA pseudoknot (Dinman, 2012a, 2012b). The slippery motif, X XXZ ZZN (where XXX and ZZZ are triplets of identical bases, and N is any nucleotide) is the mandatory element in ribosomal frameshifting (Jacks et al., 1988; Sharma et al., 2014). Pseudoknots result in pausing of the ribosome (Namy et al., 2006) and it was suggested that the stronger the mRNA pseudoknot the higher the frameshifting efficiency (Hansen et al., 2007).

Ribosomal frameshifting has been proposed to increase protein-coding capacity of organisms with small genomes (Ketteler, 2012) and gene expression regulation (Advani and Dinman, 2016; Dinman, 2012). In bacteria, release factor 2 (RF2) is encoded in 2 overlapping ORFs and the expression of RF2 gene requires PRF, where PRF occurs in nearly 90% of all the characterised bacterial genomes (Bekaert et al., 2006). However, there is an underrepresentation of ribosomal frameshifting prone sequences in protein-coding regions in *E. coli* and most of these sequences were observed in lowly expressed genes (Gurvich et al., 2003). Ribosomal frameshifting errors within lowly expressed genes would result in a low fitness cost as a small proportion of aberrant molecules will be produced from these genes. These ribosomal frameshifting prone sequences were therefore suggested to have evolved under a nearly neutral fashion (Sharma et al., 2011).

In contrast to *T. thermophilus dnaX* RNA polymerase slippage, where the novel product is shorter than the product of standard decoding, slippage-type editing is utilised for the synthesis of MxiE, a longer protein product in the pathogen *Shigella flexneri*. MxiE is a transcriptional activator that activates transcription of the type III secretion (TTS) apparatus used for infection of the epithelial cells in human leading to shigellosis (Kane et al., 2002). Studies by Penno and colleagues (2005) demonstrated that RNA polymerase slippage during transcription of *mxiE* resulted in the addition of non-templated nucleotides which fuse two overlapping ORFs, *mxiEa* and *mxiEb*. The expression of MxiE requires RNA polymerase to slip on a homopolymeric tract of 9 Ts that lies in the overlapping region of the two ORFs: 1) *mxiEa*, a 59-codon ORF containing the translational start site and 2) *mxiEb*, a 214-codon ORF encoding the DNA binding domain.

In a separate study on the ubiquitous dimeric enzyme, triosephosphate isomerase (TIM, EC 5.3.1.1), no distinct TIM was found in the hyperthermophilic bacterium *Thermotoga maritima*. However, the enzyme was found covalently linked to phosphoglycerate kinase (PGK, EC 2.7.23), forming a bifunctional, tetrameric 654 amino acid fusion protein (Schurig et al., 1995). Schurig and colleagues provided evidence that *T. maritima* produces a bifunctional enzyme with both PGK and TIM activity rather than an enzyme that only has the TIM activity. The PGK-TIM fusion protein shares the N-terminal domain of 399 amino acids with PGK and the following ~255 amino acids with TIM. These observations suggested that the homopolymeric tract prior to the N-terminal stop codon of the *pgk* gene acts as a slippage site in which RNA polymerase slippage fuses the genes together, producing a bi-functional PGK-TIM protein (Schurig et al., 1995).

The utilisation of slippage-type editing in the synthesis of additional proteins from a single gene has also been well documented in Ebola virus (EBOV) (Lee and Saphire, 2009; Sanchez et al., 1996; Volchkov et al., 1995; Volchkov et al., 2001). Intriguingly, the EBOV genome only consists of seven genes, but more than seven proteins are produced through co-transcriptional editing and post-translational processing of the glycoprotein (GP) gene. The GP gene consists of a homopolymeric tract of 7 As within a predicted hairpin loop (Volchkova et al., 1998, 1999), and a translational stop codon in the middle of the gene that prevents the synthesis of full-length GP. Most of the unedited mRNAs (approximately 80%) are directly synthesised as non-structural soluble glycoproteins (sGP) while the insertion of non-templated adenosine by slippage-type editing produces transmembrane GP. The action of RNA polymerase slipping merges the overlapping reading frames creating both the sGP and GP proteins, which share identical NH_2-terminal ends but differ in the COOH termini. In addition to the production of multiple products from a single gene, slippage-type editing has also been documented to play a role in rescuing gene expression in situations where a gene acquires a frameshift mutation, as will now be discussed.

Gene function is rescued by slippage-type editing

The presence of slippage-prone sequences in genes encoding a single functional protein product could be detrimental, but not all events of slippage are disadvantageous. Slippage has also been found to restore the reading frame of genes carrying frameshift mutations, thus rescuing gene function. Slippage-type editing has been reported to have phenotypic consequences in mammalian cells and was shown to compensate for a frameshift mutation in the same or regions of close proximity of an essential gene as demonstrated in a -1 frameshift mutation in the apolipoprotein B gene (Linton et al., 1992, 1997), factor VIII gene (Young et al., 1997) and *AP3B1* gene (Benson et al., 2004). A deletion of a C residue within an A_6CA_3

tract of the *apoB* gene was expected to result in a truncated protein but due to slippage-type editing, the reading frame was found to be restored in 10% of the transcripts, allowing the production of a full-length protein in addition to the truncated protein (Linton et al., 1992, 1997). Similar observations have been made in moderate haemophilia A, a blood disorder caused by a deletion of a single T residue in an A_8TA_2 tract of the X-linked factor VIII gene. Analysis of the DNA and RNA from patients suggests that the reading errors at the DNA level were partially restored, resulting in a milder phenotype (Young et al., 1997). Another example of reading frame correction utilising slippage-type editing was documented in canines, where an extra A was inserted within an A_9 tract of the *AP3B1* gene resulting in neutropenia (Benson et al., 2004). This mutation was predicted to cause a frameshift and create a premature stop codon but a population of *AP3B1* mRNA transcripts containing wild-type transcripts of 9As and mutant transcripts of 10As were observed. These studies suggest that frameshift defects may be partially corrected at the RNA level and ameliorate an expected severe phenotype by producing a proportion of functional AP3B1 proteins.

Slippage-type editing has also been well documented and studied in *Escherichia coli*. *In vivo* studies of transcriptional slippage in *E. coli* by Wagner et al. (1990) showed that RNA polymerase slipped at runs of 11 adenine or thymine residues at the 5' end of an out-of-frame *lacZ* gene. Surprisingly, a high level of β-galactosidase was observed in *E. coli*. suggesting that RNA polymerase slippage produced mRNAs of varying lengths where a proportion of transcripts were in the correct frame leading to the production of full-length β-galactosidase. Additionally, slippage-type editing was also experimentally demonstrated at poly(A) tracts within genes of bacterial endosymbionts by Tamas and colleagues (2008). *Buchnera aphidicola,* obligate bacterial endosymbionts of aphids, exhibit a drastic reduction in genome size (Gil et al., 2002; Ham et al., 2003), and their genomes are A+T rich (Shigenobu et al.,

2000). Tamas et al. (2008) showed that homopolymeric tracts, particularly tracts of nine or more As or Ts, are susceptible to RNA polymerase slippage and slippage results in misinterpretation of information stored in the endosymbiont genome or correction of frameshift mutations (Tamas et al., 2008). Interestingly, it was observed that RNA polymerase slippage restored and rescued gene function of frameshifted genes, but ultimately reduced the expression efficiency of in-frame genes. Even though slippage may result in reduced efficiency of gene expression of in-frame genes, it was shown that between 5 and 50% of *Buchnera* genes contain poly(A) tracts of 10 As or more (Tamas et al., 2008).

Slippage-type editing seems to offer no inherent advantage to the cell, therefore if slippage were to have evolved adaptively, one would predict the following: 1) The substrates (editable sites) for slippage-type editing would be selected for but as previously stated, polyA/T tracts that are vulnerable to slippage errors during replication and transcription are under-represented in bacterial genomes (Baranov et al., 2005; Orsi et al., 2010); 2) Gene expression efficiency would increase in the presence of slippage-type editing but as discussed, the production of heterogeneous populations of mRNAs with varying length and open reading frames essentially reduces gene expression efficiency (Tamas et al., 2008).

To summarise, slippage-type editing has been shown to be a rather robust yet low-fidelity process analogous to RNA editing. Although both processes have been speculated to evolve to play specific roles in gene expression regulation and pathways, no evolutionary evidence has been provided to support this.

Problems with studies on CNE and the evolution of editing-type processes

At this point I have reviewed discussions of neutral evolution of complex systems, where hypotheses have been inferred for a whole system based on studies of individual components of complex machinery (as discussed with the spliceosome individually, rather than a complete study of the whole spliceosome machinery). Finnigan's study, as an exception to this, is one of the more compelling, due to an experimental demonstration of complexity arising through a non-adaptive manner (Finnigan et al., 2012). However, despite this, the debate on the evolution of the spliceosome remains a hot topic, as insufficient experimental data and concrete results were provided to argue either for or against neutral evolution. While CNE has also been proposed as an explanation for the emergence of RNA editing in a number of studies (Gray et al., 2010; Stoltzfus, 1999), this concept has yet to be experimentally tested. Like the spliceosome example discussed above, experimental demonstration of non-adaptive emergence of complexity would help elucidate the plausibility of the hypotheses proposed by Gray et al., (2010) and Stoltzfus (1999). To test the hypotheses for RNA editing, in particular, we can couple tools such as experimental evolution (Travisano et al., 1995) and whole genome sequencing, to try evolve RNA editing in a system where it is not known to occur.

In experimental evolution, a subject species is often propagated for a period of time under a defined set of conditions in order to answer evolutionary questions. This combination of experimental evolution and genome sequencing has been used extensively to study viruses (Wichman et al., 2005), bacteria (Barrick et al., 2009), yeast (Lang et al., 2013) and flies (Burke et al., 2010), and provides and effective way to study the evolutionary dynamics and trajectory of model organisms. Such examples of this are seen in studies investigating underlying evolutionary processes behind the emergence of antibiotic resistance (Toprak et

al., 2011), metabolic adaptation (Wortel et al., 2016), cellular stresses and the evolution of multicellularity (Ratcliff et al., 2012) in a controlled system. Richard Lenski's group have been one of the notable pioneers of long-term evolution experiments, having evolved and tracked the genotypic and phenotypic changes of 12 initially identical populations of E.coli over 50,000 generations in a glucose-limiting media (Lenski et al., 1991). In a more recent experiment, Lenski and colleagues sequenced 264 genomes of these 12 evolving E. coli populations, and compared them to lines from a mutation accumulation (MA) experiment, where lines had been subjected to serial bottlenecking (Tenaillon et al., 2016). Mutation accumulation experiments, in comparison, maintain multiple independent and initially identical lineages at very low effective population sizes (Ne) for many generations, allowing mutations to accumulate over the generations (Halligan and Keightley, 2009). The major difference seen between the two different experimental conditions is in population structure, which ultimately results in a difference in what mutations are accumulated. Under adaptive conditions, such as those seen in the glucose experiment, selection favours beneficial mutations with deleterious mutations generally being lost. However, in a MA experiment, selection loses its discriminatory power resulting in a far higher chance of deleterious mutations being fixed in the population (Elena and Lenski, 2003), unless their fitness effects are large or lethal (Sung et al., 2012). In other words, bottlenecks imposed in a MA experiment allow mutations to accumulate at rates dependent on their underlying mutation rate, regardless of fitness effects (Barrick et al., 2009), although the selection coefficient (s, or fitness effects) of individual mutations also plays a role, with some deleterious mutations being purged due to extreme fitness effects (such as lethality) (Estes et al., 2004), therefore making it difficult to predict the frequency distribution of types of mutations (Eyre-Walker and Keightley, 2007). However, despite this, MA experiments followed by whole genome sequencing provide a useful tool to assess the role of drift in evolution, determine the rates of

spontaneous mutations, and importantly, for RNA editing, a MA experiment enables us to lessen the effects of selection and see whether RNA editing could evolve under non-adaptive conditions.

Muller's ratchet and the accumulation of mutations

Mutations that increase fitness are the basis of adaptive evolution, and most mutations with appreciable fitness effects are deleterious (Muller, 1950). Small asexual populations, such as those undergoing MA experiments, tend to gradually and irreversibly accumulate slightly deleterious mutations in a process known as Muller's ratchet (Felsenstein, 1974; Muller, 1964), while the presence of genetic recombination in large sexual populations has been shown to counteract this effect (Smith and Maynard-Smith, 1978). This fixation of slightly deleterious mutations by genetic drift occurs due to the finite size of the population, and may lead to a loss of fitter individuals in a finite population (Haigh, 1978; Metzger and Eule, 2013) (Figure 1.4). With the loss of the least mutated class, the ratchet 'clicks' leading to an irreversible increase in mutational load. The mutational load increases in a ratchet-like manner with each successive loss of the least mutated individuals leading to a gradual decline in mean fitness, which may ultimately manifest into a higher risk of population extinction (Lynch et al., 1990, 1995). MA is recognised as a threat to asexual populations and the continuous accumulation of mutations may lead to the eventual extinction of a population if: 1) The population size is small; 2) There is an absence of recombination, back and beneficial mutations; 3) Mutation rates are high and; 4) Purifying selection is limited (Gabriel et al., 1993).

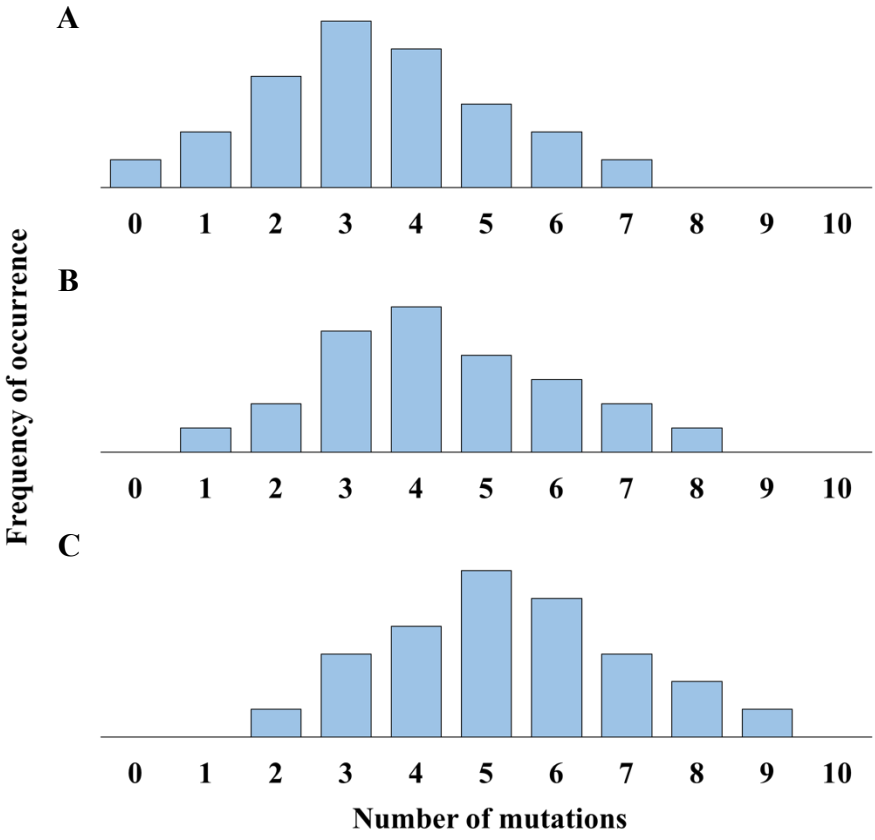

Figure 1.4 | Slightly deleterious mutations accumulate in asexual populations as a result of Muller's ratchet. A) Initially, a genotype with zero mutations exists in the population, but as mutations accumulate, this class with zero mutations (fittest) is lost over time by drift. **B)** In the absence of recombination and back mutation, the fittest, or the class with fewer mutations, cannot be restored. **C)** The whole distribution shifts to the right in a click of the ratchet and the process continues.

Muller stated that "sex is not a necessity but it is a luxury" (Muller, 1932). One of the explanations for the prevalence of sex is that it may generate genetic variation in heterogeneous, novel or changing environments, and produce some individuals with a fit combination of alleles (Smith and Maynard-Smith, 1978). Alternatively, sex could reduce or prevent the increase of mutational load, and it can affect mutation accumulation through its effect on the effective population size (N_e). In the absence of recombination, N_e is reduced,

and thus, lowers the efficacy of selection against deleterious mutations (Ota and Kimura, 1971). These theories of the evolution and prevalence of sex are encapsulated in Muller's ratchet (Felsenstein, 1974; Muller, 1964), and since its discovery, this phenomenon has been proposed to account for the advantage of sex and recombination (Chao et al., 1997; Gessler and Xu, 1999), the degeneration of the Y chromosome (Charlesworth and Charlesworth, 2000; Engelstädter, 2008; Gordo and Charlesworth, 2001), supernumerary chromosomes (Green, 1990), genome size reduction in endosymbionts (Andersson and Kurland, 1998; Mccutcheon and Moran, 2012) and mutational meltdown of small asexual populations (Gabriel et al., 1993; Lynch et al., 1993).

Muller's ratchet is most prominent in small asexual lines of descent, and the progression of Muller's ratchet has been widely reported in laboratory evolution experiments with protozoa (Bell, 1988), a DNA virus (Jaramillo et al., 2013), various classes of RNA viruses (Cervera and Elena, 2016; Chao, 1990; Chao et al., 1992; Clarke et al., 1993; Duarte et al., 1992), abiotic RNA molecules (Soll et al., 2007), bacteria (Andersson and Hughes, 1996; Tenaillon et al., 2016) and yeast (Zeyl et al., 2001). MA experiments have indisputably shown that spontaneous mutations that have deleterious effects on fitness are far more common than those that result in increases in fitness. Although it has been well-documented that the accumulation of mutations is prevalent in asexual organisms, as aforementioned, Muller's ratchet is also applicable for sexual organisms harbouring asexually propagating genomes as observed in mitochondrial tRNAs (Lynch, 1996) and neo-Y chromosomes (Kaiser and Charlesworth, 2010; Rice, 1994).

As a whole, small asexual populations are more likely to accumulate slightly deleterious mutations in an irreversible ratchet-like manner under conditions of low selection and strong

drift. In the absence of recombination, back mutations cannot stop the action of Muller's ratchet as the rate of backward and compensatory mutations are much lower than the rate of forward mutation (Haigh, 1978; Smith and Maynard-Smith, 1978). In addition, a forward mutation can occur at any site, but a backward mutation has to be at the site that reinstates the original mutation-free sequence. Therefore, the synergistic effect of an increase in the number of mutations and decrease in population size would ultimately result in the loss of fitness and mutational meltdown (Gabriel et al., 1993; Lynch et al., 1993). The detrimental effects of Muller's ratchet, however, may be counteracted by recombination and other processes that I will now discuss.

Compensation for the action of Muller's ratchet and mutational meltdown

Loss of fitness is inevitable as mutational load increases in small asexual populations due to drift, as these populations are incapable of restoring the wild-type stage by recombination. The accumulation of mutations would ultimately manifest to population extinction, a process commonly known as mutational meltdown (Gabriel et al., 1993; Lynch et al., 1993). While genetic recombination has been suggested as the principle mechanism to halt the ratchet, evolutionary theory suggests that compensatory mutations (Pfaffelhuber et al., 2012; Poon and Otto, 2000), horizontal acquisition of foreign DNA (Lorenz and Wackernagel, 1994; Overballe-Petersen and Willerslev, 2014; Overballe-Petersen et al., 2013; Takeuchi et al., 2014) and host-level selection of endosymbiont populations (Allen et al., 2009; Pettersson and Berg, 2007; Rispe et al., 2000) may also assist in decelerating the ratchet.

In animals, the streamlined mitochondrial genome, which encodes only a handful of genes, has low detectable rates of genetic recombination, yet high mutation accumulation, in most taxonomic groups (Moritz et al., 1987; Rokas et al., 2003). This high rate of the accumulation

of mutations in mitochondrial DNA (mtDNA) may have resulted from small effective population sizes associated with effectively haploid inheritance and uniparental transmission (Lynch and Blanchard, 1998). The population of animal mitochondrial genomes fits the prerequisite of Muller's ratchet: they possess both little to no genetic recombination and small effective population sizes (Kurland, 1992; Lynch, 1996, 1997), and studies have shown that animal mtDNA molecules go through genetic bottlenecks of approximately 200 mtDNA during embryonic development (Lynch, 1996; Wai et al., 2008). Additionally, mtDNA have a particularly high mutation rate (Brown et al., 1979; Denver et al., 2000; Xu et al., 2012), and its nature of low recombination rates means mutations cannot be eliminated from the population, allowing the accumulation of mutations. Thus, a mutation will persist until a stochastic process of genetic drift causes it to be lost or to dominate the entire population of mitochondrial genomes. If the accumulation of deleterious mutations is severe enough, the population could succumb to mutational meltdown, and eventually extinction (Gabriel et al., 1993). Therefore, with little to no genetic recombination, could the effect of Muller's ratchet be compensated or halted in mtDNA? In this scenario, RNA editing potentially provides an escape route as it allows the mitochondria to correct potentially deleterious mutations. Börner and colleagues (1997) suggested that various forms of RNA editing can act as 'salvage mechanisms' and proposed that RNA editing systems may have evolved to compensate for mutations arising due to Muller's ratchet in the absence of recombination. Previously I have discussed the neutral evolution of gRNA-mediated RNA editing in kinetoplastids, and I will now extend the discussion to how RNA editing may help buffer the effects of Muller's ratchet in the mitochondria of trypanosomes.

Kinetoplast DNA (kDNA), or the mitochondrial DNA of trypanosomes, generally consists of an enormous network of interlocked catenated maxicircles (that encode typical mitochondria

gene products) and minicircles (Shapiro, 1993; Sturm and Simpson, 1990) that contain genes encoding guide RNAs (gRNAs) used as templates in maxicircle RNA editing (Sturm and Simpson, 1990). RNA editing is extensive in kinetoplastids of trypanosomes, and in 12 out of 18 mRNA transcripts, uridine insertion and deletion is seen (Estévez and Simpson, 1999). In this way, RNA editing is able to correct errors in maxicircles by compensating for frameshift mutations (Simpson et al., 2000). This means that the effects of slightly deleterious mutations (the hallmark of Muller's ratchet) can be muted by the presence of RNA editing. Further to this, mutations that do persist can also be muted by the presence of multi-genome copies of both the maxicircles and minicircles and a large gRNA-coding minicircle repertoire as they provide gRNA redundancy, where the loss of a functional minicircle by mutation would not disrupt the editing process (Lukeš et al., 2002). Depending on the organism, approximately 50 maxicircles (20 to 40 kb) and more than 1,000 minicircles of different size (0.5 to 1.0 kb) and sequence heterogeneity have been reported (Simpson et al., 2000; Stuart, 1983). Mutations in one copy of a given gene will be neutral, regardless of their fitness effects, and both the mutant and unaltered copy of the gene can become fixed by drift. Alongside this, the large size variation of both minicircles and maxicircles across a range of organisms, as well as the existence of multiple copies of both maxicircles and minicircles (with large variability in the copy number of minicircles), the presence of guide RNA genes on both the minicircles and maxicircles, and variant guide RNAs with mismatches in the guide regions (Simpson et al., 2000) all show further evidence for the role of drift in editing processes, as huge variation is present and tolerated. While a number of mechanisms in kinetoplastids are compensating for the effect of Muller's ratchet, it is important to note that RNA editing is one of these mechanisms. Even with varying genotypes, phenotypes remain the same, alleviating the effects of deleterious mutations.

To conclude, Muller's ratchet is prominent in small asexual populations and can result in an increase in mutational load, loss of fitness and, ultimately, mutational meltdown. Although Muller's ratchet can be counteracted under certain circumstances, the fixation of mutations with mildly deleterious effects on fitness by genetic drift can be of great importance in the evolution of an organism. We have seen that RNA editing is capable of compensating the effects of Muller's ratchet, however, such functions can arise independently of their evolutionary origins. This means that RNA editing could have initially evolved in a non-adaptive manner, but has subsequently performed a role in a variety of different cell functions. We now aim to experimentally test the emergence of RNA editing in this thesis, by means of experimental evolution, to assess whether RNA editing may evolve in a non-adaptive manner and what effect this may have on gene expression and cell fitness.

Summary

It is plausible that complex editing-type processes may have evolved non-adaptively however, there is no clear experimental evidence to this reasoning. Covello and Gray (1993) proposed a 3-step neutral model for the evolution of RNA editing which could be generalized to explain the evolution of slippage-type editing (Box 1.1). Under this model, the editing machinery, slippage-prone RNA polymerase, exists before there was a need for editing. Subsequently, frameshift mutations occur at editable sites (long tract of As or Ts) and may be fixed in the population by drift, leading to reduced efficiency of gene expression. Upon fixation of these mutations, slippage-type editing is now indispensable for gene expression and is therefore maintained by selection.

Although this phenomenon has been described, and a role for drift in driving its emergence has been proposed, this has not been explicitly tested. The primary aim of this study was to test the role of genetic drift on the evolution of complex processes such as slippage-type editing. While *E. coli* RNA polymerase is slippage-prone, there is an underrepresentation of homopolymeric tracts in free-living *E. coli* genomes (Baranov et al., 2005; Orsi et al., 2010). Studies on population genetics predict that selection is weaker and drift is stronger in small bottlenecked populations, and this have previously been observed in *Buchnera* (Moran, 1996). As previously discussed, *Buchnera* are endosymbionts of aphids, which go through population bottlenecks through each maternal transmission, and homopolymeric tracts are abundant in their genomes. Therefore, we subjected *E. coli* populations to serial single-colony bottlenecks, mimicking the population genetic structure of *Buchnera*. This will enable us to directly test the role of genetic drift in driving the emergence of complex processes such as editing. Additionally, in small asexual populations we might see the fixation of slightly deleterious mutations due to the absence of recombination and strong selection. We might expect to see a potential loss of fitness and increase in mutational load.

In Chapter 2, we examined the phenomenon of Muller's ratchet in *E. coli* populations that have undergone serial single-colony bottlenecks by assessing the effect of the ratchet at the genotypic and phenotypic levels. Although mutation accumulation (MA) and the effect of Muller's ratchet have been extensively studied, in this thesis we utilised a novel method to identify functional variations between ancestor wild-type *E. coli* genomes and bottlenecked genomes using HMM profile-based methods (Wheeler et al., 2016). Briefly, we predicted the effects of mutations on protein functionality by first calculating bitscore values of the ancestor and evolved bottlenecked protein sequences based on the sequence alignment to HMM profiles built from gamma-proteobacterial protein sequences. Subsequently, by

subtracting the bitscore of the ancestor protein (reference) from that of the bottleneck protein (variant), a measure of divergence between proteins is produced, termed Delta-bitscore (DBS). The calculated DBS values allow us to predict whether the mutations would result in a gain or loss of function in a protein.

In Chapter 3, we tested Covello and Gray's (1993) model for the evolution of RNA editing, and provide evidence for the neutral emergence of complex slippage-type editing in *E. coli*. Under conditions of strong drift and low selection, we observed the emergence of frameshifted slippage-prone homopolymeric tracts that require editing for the production of full-length proteins. We then assessed the effect of frameshifted polyT in *araC* (observed in our evolution experiment) on cell fitness by comparing growth rates of wild-type and frameshifted polyT *araC* knock-in in arabinose only media. Under this condition where arabinose is the sole carbon source, functional AraC is essential for cell survival and RNA polymerase slippage is required to produce full-length AraC in the knock-in strain.

In Chapter 4, we examined the impact of slippage-type editing on gene expression and protein production. Although the phenomenon of slippage-type editing has been well documented in *Buchnera* and its effect on gene expression has been demonstrated, there are no reports on the impact of slippage on protein production. Using GFP reporter systems, we assessed the impact of slippage on protein production by measuring the relative fluorescence units (RFU) and by Western blotting. To conclude, we tie this work together and discuss how our results can aid in explaining the origins of RNA editing in eukaryotes.

References

Abad, M.G., Long, Y., Willcox, A., Gott, J.M., Gray, M.W., and Jackman, J.E. (2011). A role for tRNA(His) guanylyltransferase (Thg1)-like proteins from *Dictyostelium discoideum* in mitochondrial 5'-tRNA editing. RNA *17*, 613–623.

Allen, J.M., Light, J.E., Perotti, M.A., Braig, H.R., and Reed, D.L. (2009). Mutational meltdown in primary endosymbionts: selection limits Muller's ratchet. PLoS ONE *4*, e4969.

Andersson, D.I., and Hughes, D. (1996). Muller's ratchet decreases fitness of a DNA-based microbe. Proc. Natl. Acad. Sci. *93*, 906–907.

Andersson, S.G.., and Kurland, C.G. (1998). Reductive evolution of resident genomes. Trends Microbiol. *6*, 263–268.

Anikin, M., Molodtsov, V., Temiakov, D., and McAllister, W.T. (2010). Transcript slippage and recoding. In Recoding: Expansion of Decoding Rules Enriches Gene Expression, J.F. Atkins, and R.F. Gesteland, eds. (Springer New York), pp. 409–432.

Aphasizhev, R., and Aphasizheva, I. (2011), Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer. WIREs RNA, *2*: 669–685. doi:10.1002/wrna.82

Aphasizhev, R., Sbicego, S., Peris, M., Jang, S.-H., Aphasizheva, I., Simpson, A.M., Rivlin, A., and Simpson, L. (2002). Trypanosome mitochondrial 3′ terminal uridylyl transferase (TUTase): the key enzyme in U-insertion/deletion RNA editing. Cell *108*, 637–648.

Baranov, P.V., Hammer, A.W., Zhou, J., Gesteland, R.F., and Atkins, J.F. (2005). Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. Genome Biol. *6*, R25.

Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E., and Kim, J.F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature *461*, 1243–1247.

Bell, G. (1988). Sex and Death in Protozoa: The History of Obsession (Cambridge University Press).

Benne, R., Van Den Burg, J., Brakenhoff, J., P.J., Sloof, P., Van Boom, J., H., and Tromp, M., C. (1986). Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell *46*, 819–826.

Benson, K.F., Person, R.E., Li, FQ., Williams, K., and Horwitz, M. (2004). Paradoxical homozygous expression from heterozygotes and heterozygous expression from homozygotes as a consequence of transcriptional infidelity through a polyadenine tract in the AP3B1 gene responsible for canine cyclic neutropenia. Nucleic Acids Res. *32*, 6327–6333.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science *277*, 1453–1462.

Blum, B., Bakalara, N., and Simpson, L. (1990). A model for RNA editing in kinetoplastid mitochondria: “guide” RNA molecules transcribed from maxicircle DNA provide the edited information. Cell *60*, 189–198.

Bonner, J.T. (1988). The Evolution of Complexity by Means of Natural Selection (Princeton, N.J: Princeton University Press).

Börner, G.V., Yokobori, S., Mörl, M., Dörner, M., and Pääbo, S. (1997). RNA editing in metazoan mitochondria: staying fit without sex. FEBS Lett. *409*, 320–324.

Brennicke, A., Marchfelder, A., and Binder, S. (1999). RNA editing. FEMS Microbiol. Rev. *23*, 297–316.

Brown, W.M., George, M., and Wilson, A.C. (1979). Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. *76*, 1967–1971.

Bullerwell, C.E., and Gray, M.W. (2005). In vitro characterization of a tRNA editing activity in the mitochondria of *Spizellomyces punctatus*, a Chytridiomycete fungus. J. Biol. Chem. *280*, 2463–2470.

Burger, G., Yan, Y., Javadi, P., and Lang, B.F. (2009). Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals. Trends Genet. *25*, 381–386.

Burke, M.K., Dunham, J.P., Shahrestani, P., Thornton, K.R., Rose, M.R., and Long, A.D. (2010). Genome-wide analysis of a long-term evolution experiment with Drosophila. Nature *467*, 587–590.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. Trends Genet. *7*, 145–148.

Cavalier-Smith, T. (1997). Cell and genome coevolution: facultative anaerobiosis, glycosomes and kinetoplastan RNA editing. Trends Genet. *13*, 6–9.

Cech, T.R. (1986). The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell *44*, 207–210.

Cervera, H., and Elena, S.F. (2016). Genetic variation in fitness within a clonal population of a plant RNA virus. Virus Evol. *2*, vew006.

Chamberlin, M., and Berg, P. (1962). Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from *Escherichia coli*. Proc. Natl. Acad. Sci. *48*, 81–94.

Chao, L. (1990). Fitness of RNA virus decreased by Muller’s ratchet. Nature *348*, 454–455.

Chao, L., Tran, T., and Matthews, C. (1992). Muller’s ratchet and the advantage of sex in the RNA virus. Evolution *46*, 289–299.

Chao, L., Tran, T.T., and Tran, T.T. (1997). The Advantage of Sex in the RNA Virus φ6. Genetics *147*, 953–959.

Charlesworth, B., and Charlesworth, D. (2000). The degeneration of Y chromosomes. Philos. Trans. R. Soc. B Biol. Sci. *355*, 1563–1572.

Clarke, D.K., Duarte, E.A., Moya, A., Elena, S.F., Domingo, E., and Holland, J. (1993). Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses. J. Virol. *67*, 222–228.

Copertino, D.W., and Hallick, R.B. (1993). Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. Trends Biochem. Sci. *18*, 467–471.

Covello, P.S., and Gray, M.W. (1989). RNA editing in plant mitochondria. Nature *341*, 662–666.

Covello, P.S., and Gray, M.W. (1993). On the evolution of RNA editing. Trends Genet. *9*, 265–268.

Decker, C.J., and Sollner-Webb, B. (1990). RNA editing involves indiscriminate U changes throughout precisely defined editing domains. Cell *61*, 1001–1011.

Deng, L., and Shuman, S. (1997). Elongation properties of vaccinia virus RNA polymerase: pausing, slippage, 3' end addition, and termination site choice. Biochemistry (Mosc.) *36*, 15892–15899.

Denver, D.R., Morris, K., Lynch, M., Vassilieva, L.L., and Thomas, W.K. (2000). High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. Science *289*, 2342–2344.

Deschamps, P., Lara, E., Marande, W., López-García, P., Ekelund, F., and Moreira, D. (2011). Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. Mol. Biol. Evol. *28*, 53–58.

Duarte, E., Clarke, D., Moya, A., Domingo, E., and Holland, J. (1992). Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc. Natl. Acad. Sci. *89*, 6015–6019.

Elena, S.F., and Lenski, R.E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat. Rev. Genet. *4*, 457–469.

Engelstädter, J. (2008). Muller's ratchet and the degeneration of Y chromosomes: a simulation study. Genetics *180*, 957–967.

Estes, S., Phillips, P.C., Denver, D.R., Thomas, W.K., and Lynch, M. (2004). Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in *Caenorhabditis elegans*. Genetics *166*, 1269–1279.

Estévez, A.M., and Simpson, L. (1999). Uridine insertion/deletion RNA editing in trypanosome mitochondria - a review. Gene *240*, 247–260.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat. Rev. Genet. *8*, 610–618.

Felsenstein, J. (1974). The evolutionary advantage of recombination. Genetics *78*, 737–756.

Fernández, A., and Lynch, M. (2011). Non-adaptive origins of interactome complexity. Nature *474*, 502–505.

Finnigan, G.C., Hanson-Smith, V., Stevens, T.H., and Thornton, J.W. (2012). Evolution of increased complexity in a molecular machine. Nature *481*, 360–364.

Flower, A.M., and McHenry, C.S. (1990). The gamma subunit of DNA polymerase III holoenzyme of *Escherichia coli* is produced by ribosomal frameshifting. Proc. Natl. Acad. Sci. *87*, 3713–3717.

Foote, M., and Gould, S.J. (1992). Cambrian and recent morphological disparity. Science *258*, 1816.

Forgac, M. (2007). Vacuolar ATPases: rotary proton pumps in physiology and pathophysiology. Nat. Rev. Mol. Cell Biol. *8*, 917–929.

Forget, L., Ustinova, J., Wang, Z., Huss, V.A.R., and Lang, B.F. (2002). *Hyaloraphidium curvatum*: a linear mitochondrial genome, tRNA editing, and an evolutionary link to lower fungi. Mol. Biol. Evol. *19*, 310–319.

Fortunato, S.A.V., Adamski, M., Ramos, O.M., Leininger, S., Liu, J., Ferrier, D.E.K., and Adamska, M. (2014). Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature *514*, 620–623.

Gabriel, W., Lynch, M., and Burger, R. (1993). Muller's ratchet and mutational meltdowns. Evolution *47*, 1744–1757.

Gessler, D.D., and Xu, S. (1999). On the evolution of recombination and meiosis. Genet. Res. *73*, 119–131.

Gil, R., Sabater-Muñoz, B., Latorre, A., Silva, F.J., and Moya, A. (2002). Extreme genome reduction in Buchnera spp.: Toward the minimal genome needed for symbiotic life. Proc. Natl. Acad. Sci. *99*, 4454–4458.

Gilbert, W. (1986). Origin of life: the RNA world. Nature *319*, 618–618.

Gordo, I., and Charlesworth, B. (2001). The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes. Genet. Res. *78*, 149–161.

Gould, S.J. (1997). Full House: The Spread of Excellence from Plato to Darwin (Three Rivers Press).

Gould, S.J., and Lewontin, R.C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc. R. Soc. Lond. B Biol. Sci. *205*, 581–598.

Gray, M. (2003). Diversity and Evolution of Mitochondrial RNA Editing Systems. IUBMB Life *55*, 227–233.

Gray, M.W. (2012a). Mitochondrial evolution. Cold Spring Harb. Perspect. Biol. *4*, a011403.

Gray, M.W. (2012b). Evolutionary origin of RNA editing. Biochemistry (Mosc.) *51*, 5235–5242.

Gray, M.W. (2013). RNA editing: evolutionary implications. In eLS, John Wiley & Sons, Ltd, ed. (Chichester, UK: John Wiley & Sons, Ltd).

Gray, M.W., and Covello, P.S. (1993). RNA editing in plant mitochondria and chloroplasts. FASEB J. *7*, 64–71.

Gray, M.W., and Doolittle, W.F. (1982). Has the endosymbiont hypothesis been proven? Microbiol. Rev. *46*, 1–42.

Gray, M.W., Burger, G., and Lang, B.F. (1999). Mitochondrial evolution. Science *283*, 1476–1481.

Gray, M.W., Lukeš, J., Archibald, J.M., Keeling, P.J., and Doolittle, W.F. (2010). Irremediable complexity? Science *330*, 920–921.

Green, D.M. (1990). Muller's ratchet and the evolution of supernumerary chromosomes. Genome *33*, 818–824.

Haigh, J. (1978). The accumulation of deleterious genes in a population - Muller's ratchet. Theor. Popul. Biol. *14*, 251–267.

Hajduk, S.L., Harris, M.E., and Pollard, V.W. (1993). RNA editing in kinetoplastid mitochondria. FASEB J. *7*, 54–63.

Halligan, D.L., and Keightley, P.D. (2009). Spontaneous mutation accumulation studies in evolutionary genetics. Annu. Rev. Ecol. Evol. Syst. *40*, 151–172.

Ham, R.C.H.J. van, Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J.M., Jiménez, L., Postigo, M., Silva, F.J., et al. (2003). Reductive genome evolution in *Buchnera aphidicola*. Proc. Natl. Acad. Sci. *100*, 581–586.

Heinemann, I.U., Randau, L., Tomko, R.J., and Söll, D. (2010). 3′–5′ tRNAHis guanylyltransferase in bacteria. FEBS Lett. *584*, 3567–3572.

Hetzer, M., Wurzer, G., Schweyen, R.J., and Mueller, M.W. (1997). Trans-activation of group II intron splicing by nuclear U5 snRNA. Nature *386*, 417–420.

Hiesel, R., Wissinger, B., Schuster, W., and Brennicke, A. (1989). RNA editing in plant mitochondria. Science *246*, 1632–1634.

Jacob, F. (1977). Evolution and tinkering. Science *196*, 1161–1166.

Jaramillo, N., Domingo, E., Muñoz-Egea, M.C., Tabarés, E., and Gadea, I. (2013). Evidence of Muller's ratchet in herpes simplex virus type 1. J. Gen. Virol. *94*, 366–375.

Kaiser, V.B., and Charlesworth, B. (2010). Muller's ratchet and the degeneration of the *Drosophila miranda* neo-Y chromosome. Genetics *185*, 339–348.

Kane, C.D., Schuch, R., Day, W.A., and Maurelli, A.T. (2002). MxiE regulates intracellular expression of factors secreted by the *Shigella flexneri* 2a type III secretion system. J. Bacteriol. *184*, 4409–4419.

Kugita, M., Yamamoto, Y., Fujikawa, T., Matsumoto, T., and Yoshinaga, K. (2003). RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res. *31*, 2417–2423.

Kurland, C.G. (1992). Evolution of mitochondrial genomes and the genetic code. BioEssays News Rev. Mol. Cell. Dev. Biol. *14*, 709–714.

Laforest, M.J., Roewer, I., and Lang, B.F. (1997). Mitochondrial tRNAs in the lower fungus *Spizellomyces punctatus*: tRNA editing and UAG "stop" codons recognized as leucine. Nucleic Acids Res. *25*, 626–632.

Landweber, L.F., and Gilbert, W. (1993). RNA editing as a source of genetic variation. Nature *363*, 179–182.

Lang, G.I., Rice, D.P., Hickman, M.J., Sodergren, E., Weinstock, G.M., Botstein, D., and Desai, M.M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature *500*, 571–574.

Larsen, B., Wills, N.M., Nelson, C., Atkins, J.F., and Gesteland, R.F. (2000). Nonlinearity in genetic decoding: homologous DNA replicase genes use alternatives of transcriptional slippage or translational frameshifting. Proc. Natl. Acad. Sci. *97*, 1683–1688.

Lee, J.E., and Saphire, E.O. (2009). Ebolavirus glycoprotein structure and mechanism of entry. Future Virol. *4*, 621–635.

Lenski, R.E. (2017). What is adaptation by natural selection? Perspectives of an experimental microbiologist. PLoS Genet. *13,* e1006668.

Lenski, R.E., Rose, M.R., Simpson, S.C., and Tadler, S.C. (1991). Long-term experimental evolution in *Escherichia coli*. I. adaptation and divergence during 2,000 generations. Am. Nat. *138*, 1315–1341.

Lenski, R.E., Ofria, C., Pennock, R.T., and Adami, C. (2003). The evolutionary origin of complex features. Nature *423*, 139–144.

Levinson, G., and Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. *4*, 203–221.

Lewontin, R.C. (2002). Directions in evolutionary biology. Annu. Rev. Genet. *36*, 1–18.

Lin, S., Zhang, H., Spencer, D.F., Norman, J.E., and Gray, M.W. (2002). Widespread and extensive editing of mitochondrial mRNAs in dinoflagellates. J. Mol. Biol. *320*, 727–739.

Linton, M.F., Pierotti, V., and Young, S.G. (1992). Reading-frame restoration with an apolipoprotein B gene frameshift mutation. Proc. Natl. Acad. Sci. U. S. A. *89*, 11431–11435.

Linton, M.F., Raabe, M., Pierotti, V., and Young, S.G. (1997). Reading-frame restoration by transcriptional slippage at long stretches of adenine residues in mammalian cells. J. Biol. Chem. *272*, 14127–14132.

Liu, C., Heath, L.S., and Turnbough, C.L. (1994). Regulation of pyrBI operon expression in *Escherichia coli* by UTP-sensitive reiterative RNA synthesis during transcriptional initiation. Genes Dev. *8*, 2904–2912.

Lonergan, K.M., and Gray, M.W. (1993). Editing of transfer RNAs in *Acanthamoeba castellanii* mitochondria. Science *259*, 812–816.

Lorenz, M.G., and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. *58*, 563–602.

Lukeš, J., Guilbride, D.L., Votýpka, J., Zíková, A., Benne, R., and Englund, P.T. (2002). Kinetoplast DNA network: evolution of an improbable structure. Eukaryot. Cell *1*, 495–502.

Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F., and Gray, M.W. (2011). How a neutral evolutionary ratchet can build cellular complexity. IUBMB Life *63*, 528–537.

Lynch, M. (1996). Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. Mol. Biol. Evol. *13*, 209–220.

Lynch, M. (1997). Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. Mol. Biol. Evol. *14*, 914–925.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. *60*, 327–349.

Lynch, M. (2007a). The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. U. S. A. *104 Suppl 1*, 8597–8604.

Lynch, M. (2007b). The Origins of Genome Architecture (Sinauer Associates).

Lynch, M., and Blanchard, J.L. (1998). Deleterious mutation accumulation in organelle genomes. Genetica *102–103*, 29–39.

Lynch, M., Gabriel, W., and Nov, N. (1990). Mutation load and the survival of small populations. Evolution *44*, 1725–1737.

Lynch, M., Bürger, R., Butcher, D., and Gabriel, W. (1993). The mutational meltdown in asexual populations. J. Hered. *84*, 339–344.

Lynch, M., Conery, J., and Burger, R. (1995). Mutational meltdowns in sexual populations. Evolution *49*, 1067.

Maniatis, T., and Reed, R. (2002). An extensive network of coupling among gene expression machines. Nature *416*, 499–506.

Massey, S.E. (2015). Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. Life Basel Switz. *5*, 1301–1332.

Mccutcheon, J.P., and Moran, N.A. (2012). Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. *10*, 13–26.

McShea, D.W. (1996). Perspective: metazoan complexity and evolution: is there a trend? Evolution *50*, 477–492.

Mendivil Ramos, O., Barker, D., and Ferrier, D.E.K. (2012). Ghost loci imply Hox and ParaHox existence in the last common ancestor of animals. Curr. Biol. CB *22*, 1951–1956.

Metzger, J.J., and Eule, S. (2013). Distribution of the fittest individuals and the rate of Muller's ratchet in a model with overlapping generations. PLoS Comput. Biol. *9*, e1003303.

Moran, N.A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. *93*, 2873–2878.

Moran, N.A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. Cell *108*, 583–586.

Moreira, D., Le Guyader, H., and Philippe, H. (2000). The origin of red algae and the evolution of chloroplasts. Nature *405*, 69–72.

Moritz, C., Dowling, T.E., and Brown, W.M. (1987). Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annu. Rev. Ecol. Syst. *18*, 269–292.

Muller, H.J. (1932). Some Genetic Aspects of Sex. Am. Nat. *66*, 118–138.

Muller, H.J. (1950). Our load of mutations. Am. J. Hum. Genet. *2*, 111–176.

Muller, H.J. (1964). The relation of recombination to mutational advance. Mutat. Res. *1*, 2–9.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. Annu. Rev. Biochem. *79*, 321–349.

Nudler, E., Mustaev, A., Goldfarb, A., and Lukhtanov, E. (1997). The RNA–DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. Cell *89*, 33–41.

O'Malley, M.A., Wideman, J.G., and Ruiz-Trillo, I. (2016). Losing complexity: the role of simplification in macroevolution. Trends Ecol. Evol. *31*, 608–621.

Orsi, R.H., Bowen, B.M., and Wiedmann, M. (2010). Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. BMC Genomics *11*, 102.

Osawa, S., and Jukes, T.H. (1989). Codon reassignment (codon capture) in evolution. J. Mol. Evol. *28*, 271–278.

Osawa, S., Jukes, T.H., Watanabe, K., and Muto, A. (1992). Recent evidence for evolution of the genetic code. Microbiol. Rev. *56*, 229–264.

Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H.-Y., Wei, W., Suzuki, S., Arashida, R., Nakata, D., Miyata, S., Ugaki, M., et al. (2004). Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. Nat. Genet. *36*, 27–29.

Ota, T., and Kimura, M. (1971). On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. *1*, 18–25.

Overballe-Petersen, S., and Willerslev, E. (2014). Horizontal transfer of short and degraded DNA has evolutionary implications for microbes and eukaryotic sexual reproduction. BioEssays *36*, 1005–1010.

Overballe-Petersen, S., Harms, K., Orlando, L.A.A., Mayar, J.V.M., Rasmussen, S., Dahl, T.W., Rosing, M.T., Poole, A.M., Sicheritz-Ponten, T., Brunak, S., et al. (2013). Bacterial natural transformation by highly fragmented and damaged DNA. Proc. Natl. Acad. Sci. *110*, 19860–19865.

Parks, A.R., Court, C., Lubkowska, L., Jin, D.J., Kashlev, M., and Court, D.L. (2014). Bacteriophage λ N protein inhibits transcription slippage by *Escherichia coli* RNA polymerase. Nucleic Acids Res. *42*, 5823–5829.

Penno, C., Sansonetti, P., and Parsot, C. (2005). Frameshifting by transcriptional slippage is involved in production of MxiE, the transcription activator regulated by the activity of the type III secretion apparatus in Shigella flexneri. Mol. Microbiol. *56*, 204–214.

Penno, C., Sharma, V., Coakley, A., O'Connell Motherway, M., van Sinderen, D., Lubkowska, L., Kireeva, M.L., Kashlev, M., Baranov, P.V., and Atkins, J.F. (2015). Productive mRNA stem loop-mediated transcriptional slippage: crucial features in common with intrinsic terminators. Proc. Natl. Acad. Sci. *112*, E1984–E1993.

Pettersson, M.E., and Berg, O.G. (2007). Muller's ratchet in symbiont populations. Genetica *130*, 199–211.

Pfaffelhuber, P., Staab, P.R., and Wakolbinger, A. (2012). Muller's ratchet with compensatory mutations. Ann. Appl. Probab. *22*, 2108–2132.

Poole, A.M., Phillips, M.J., and Penny, D. (2003). Prokaryote and eukaryote evolvability. Biosystems *69*, 163–185.

Poon, A., and Otto, S.P. (2000). Compensating for our load of mutations: freezing the meltdown of small populations. Evolution *54*, 1467–1479.

Popper, K. (2002). Conjectures and Refutations: The Growth of Scientific Knowledge (London ; New York: Routledge).

Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J., and Scott, J. (1987). A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell *50*, 831–840.

Price, D.H., and Gray, M.W. (1999). Confirmation of predicted edits and demonstration of unpredicted edits in *Acanthamoeba castellanii* mitochondrial tRNAs. Curr. Genet. *35*, 23–29.

Pullirsch, D., and Jantsch, M.F. (2010). Proteome diversification by adenosine to inosine RNA editing. RNA Biol. *7*, 205–212.

Qi, F., and Turnbough Jr, C.L. (1995). Regulation of codBA operon expression in *Escherichia coli* by UTP-dependent reiterative transcription and UTP-sensitive transcriptional start site switching. J. Mol. Biol. *254*, 552–565.

Randau, L., Stanley, B.J., Kohlway, A., Mechta, S., Xiong, Y., and Söll, D. (2009). A cytidine deaminase edits C to U in transfer RNAs in Archaea. Science *324*, 657–659.

Rao, B.S., Maris, E.L., and Jackman, J.E. (2011). tRNA 5′-end repair activities of tRNAHis guanylyltransferase (Thg1)-like proteins from Bacteria and Archaea. Nucleic Acids Res. *39*, 1833–1842.

Ratcliff, W.C., Denison, R.F., Borrello, M., and Travisano, M. (2012). Experimental evolution of multicellularity. Proc. Natl. Acad. Sci. *109*, 1595–1600.

Rice, W.R. (1994). Degeneration of a nonrecombining chromosome. Science *263*, 230–232.

Rispe, C., Moran, N.A., and Otto, A.E.S.P. (2000). Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. Am. Nat. *156*, 425–441.

Rokas, A., Ladoukakis, E., and Zouros, E. (2003). Animal mitochondrial DNA recombination revisited. Trends Ecol. Evol. *18*, 411–417.

Sanchez, A., Trappier, S.G., Mahy, B.W., Peters, C.J., and Nichol, S.T. (1996). The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proc. Natl. Acad. Sci. *93*, 3602–3607.

Schurig, H., Beaucamp, N., Ostendorp, R., Jaenicke, R., Adler, E., and Knowles, J.R. (1995). Phosphoglycerate kinase and triosephosphate isomerase from the hyperthermophilic bacterium *Thermotoga maritima* form a covalent bifunctional enzyme complex. EMBO J. *14*, 442–451.

Shapiro, T.A. (1993). Kinetoplast DNA maxicircles: networks within networks. Proc. Natl. Acad. Sci. *90*, 7809–7813.

Sharp, P.A. (1985). On the origin of RNA splicing and introns. Cell *42*, 397–400.

Shaw, J.M., Feagin, J.E., Stuart, K., and Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell *53*, 401–411.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature *407*, 81–86.

Shukla, G.C., and Padgett, R.A. (2002). A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome. Mol. Cell *9*, 1145–1150.

Simpson, L., Thiemann, O.H., Savill, N.J., Alfonzo, J.D., and Maslov, D.A. (2000). Evolution of RNA editing in trypanosome mitochondria. Proc. Natl. Acad. Sci. *97*, 6986–6993.

Simpson, L., Sbicego, S., and Aphasizhev, R. (2003). Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business. RNA *9*, 265–276.

Smith, J.M., and Maynard-Smith, J. (1978). The Evolution of Sex (Cambridge Univ Press).

Smith, J.M., and Szathmáry, E. (1995). The Major Transitions in Evolution (Oxford; New York: Oxford University Press).

Soll, S.J., Díaz Arenas, C., and Lehman, N. (2007). Accumulation of deleterious mutations in small abiotic populations of RNA. Genetics *175*, 267–275.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. J. Mol. Evol. *49*, 169–181.

Stoltzfus, A. (2012). Constructive neutral evolution: exploring evolutionary theory's curious disconnect. Biol. Direct *7*, 35.

Stuart, K. (1983). Kinetoplast DNA, mitochondria DNA with a difference. Mol. Biochem. Parasitol. *9*, 93–104.

Stuart, K. (1991). RNA editing in trypanosomatid mitochondria. Annu. Rev. Microbiol. *45*, 327–344.

Stuart, K., Allen, T.E., Heidmann, S., and Seiwert, S.D. (1997). RNA editing in kinetoplastid protozoa. Microbiol. Mol. Biol. Rev. *61*, 105–120.

Sturm, N.R., and Simpson, L. (1990). Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. Cell *61*, 879–884.

Sung, W., Ackerman, M.S., Miller, S.F., Doak, T.G., and Lynch, M. (2012). Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. *109*, 18488–18492.

Szathmáry, E., and Smith, J.M. (1995). The major evolutionary transitions. Nature *374*, 227–232.

Takeuchi, N., Kaneko, K., and Koonin, E.V. (2014). Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. G3: Genes|Genomes|Genetics *4*, 325–339.

Tamas, I., Wernegreen, J.J., Nystedt, B., Kauppinen, S.N., Darby, A.C., Gomez-Valero, L., Lundin, D., Poole, A.M., and Andersson, S.G.E. (2008). Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. Proc. Natl. Acad. Sci. *105*, 14934–14939.

Tenaillon, O., Barrick, J.E., Ribeck, N., Deatherage, D.E., Blanchard, J.L., Dasgupta, A., Wu, G.C., Wielgoss, S., Cruveiller, S., Médigue, C., et al. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. Nature *536*, 165–170.

Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D.L., and Kishony, R. (2011). Evolutionary paths to antibiotic resistance under dynamically sustained drug stress. Nat. Genet. *44*, 101–105.

Travisano, M., Mongold, J.A., Bennett, A.F., and Lenski, R.E. (1995). Experimental tests of the roles of adaptation, chance, and history in evolution. Science *267*, 87–90.

Tu, A.H., and Turnbough, C.L. (1997). Regulation of upp expression in *Escherichia coli* by UTP-sensitive selection of transcriptional start sites coupled with UTP-dependent reiterative transcription. J. Bacteriol. *179*, 6665–6673.

Turnbough, C.L. (2011). Regulation of gene expression by reiterative transcription. Curr. Opin. Microbiol. *14*, 142–147.

Uptain, S.M., Kane, C.M., and Chamberlin, M.J. (1997). Basic mechanisms of transcript elongation and its regulation. Annu. Rev. Biochem. *66*, 117–172.

Visomirski-Robic, L.M., and Gott, J.M. (1995). Accurate and efficient insertional RNA editing in isolated Physarum mitochondria. RNA *1*, 681–691.

Volchkov, V.E., Becker, S., Volchkova, V.A., Ternovoj, V.A., Kotov, A.N., Netesov, S.V., and Klenk, H.D. (1995). GP mRNA of Ebola virus is edited by the Ebola virus polymerase and by T7 and vaccinia virus polymerases. Virology *214*, 421–430.

Volchkov, V.E., Volchkova, V.A., Muhlberger, E., Kolesnikova, L.V., Weik, M., Dolnik, O., and Klenk, H.D. (2001). Recovery of infectious Ebola virus from complementary DNA: RNA editing of the GP gene and viral cytotoxicity. Science *291*, 1965–1969.

Volchkova, V.A., Feldmann, H., Klenk, H.D., and Volchkov, V.E. (1998). The nonstructural small glycoprotein sGP of Ebola virus is secreted as an antiparallel-orientated homodimer. Virology *250*, 408–414.

Volchkova, V.A., Klenk, H.D., and Volchkov, V.E. (1999). Delta-peptide is the carboxy-terminal cleavage fragment of the nonstructural small glycoprotein sGP of Ebola virus. Virology *265*, 164–171.

Wagner, L.A., Weiss, R.B., Driscoll, R., Dunn, D.S., and Gesteland, R.F. (1990). Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli*. Nucleic Acids Res. *18*, 3529–3535.

Wai, T., Teoli, D., and Shoubridge, E.A. (2008). The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. Nat. Genet. *40*, 1484–1488.

Wheeler, N.E., Barquist, L., Kingsley, R.A., and Gardner, P.P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. Bioinforma. Oxf. Engl.

Wichman, H.A., Millstein, J., and Bull, J.J. (2005). Adaptive molecular evolution for 13,000 phage generations. Genetics *170*, 19–31.

Will, C.L., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harb. Perspect. Biol. *3*, a003707.

Wortel, M.T., Bosdriesz, E., Teusink, B., and Bruggeman, F.J. (2016). Evolutionary pressures on microbial metabolic strategies in the chemostat. Sci. Rep. *6*, 29503.

Xu, S., Schaack, S., Seyfert, A., Choi, E., Lynch, M., and Cristescu, M.E. (2012). High mutation rates in the mitochondrial genomes of *Daphnia pulex*. Mol. Biol. Evol. *29*, 763–769.

Young, M., Inaba, H., Hoyer, L.W., Higuchi, M., Kazazian, H.H., and Antonarakis, S.E. (1997). Partial correction of a severe molecular defect in hemophilia A, because of errors during expression of the factor VIII gene. Am. J. Hum. Genet. *60*, 565–573.

Zauner, S., Greilinger, D., Laatsch, T., Kowallik, K.V., and Maier, U.-G. (2004). Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate Ceratium horridum. FEBS Lett. *577*, 535–538.

Zhang, J. (2010) Positive Darwinian selection in gene evolution. In: *Darwin's Heritage Today: Proceedings of the Darwin 200 Beijing International Conference*. M. Long, et al., eds, High Education Press, Beijing, Pp. 288-309.

Zeyl, C., Mizesko, M., and de Visser, J.A. (2001). Mutational meltdown in laboratory yeast populations. Evol. Int. J. Org. Evol. *55*, 909–917.

Zhou, Y.N., Lubkowska, L., Hui, M., Court, C., Chen, S., Court, D.L., Strathern, J., Jin, D.J., and Kashlev, M. (2013). Isolation and characterization of RNA polymerase rpoB mutations that alter transcription slippage during elongation in *Escherichia coli*. J. Biol. Chem. *288*, 2700–2710.

CHAPTER 2

Drift drives the accumulation of mutations and the extinction of small populations

Introduction

Mutation accumulation (MA) experiments are an effective way to assess the effects of genetic drift in driving evolution. MA experiments involve periodically bottlenecking a population, effectively reducing the effective population size (N_e), such that evolution proceeds by close to pure genetic drift (Halligan and Keightley, 2009). A number of MA experiments have been carried out in protozoa (Bell, 1988), RNA and DNA viruses (Cervera and Elena, 2016; Chao, 1990; Jaramillo et al., 2013), bacteria (Andersson and Hughes, 1996), abiotic RNA molecules (Soll et al., 2007), as well as yeast (Zeyl et al., 2001) and accumulation of mutations were consistently observed across all the experiment. The accumulation of slightly deleterious mutations is a hallmark of Muller's ratchet and is prevalant in small asexual populations undergoing population bottlenecks (Felsenstein, 1974; Muller, 1964). Under conditions of low selection and high drift, deleterious mutations are fixed in the population as a consequence of the recurrent loss of the fittest class by chance (Charlesworth and Charlesworth, 1997).

Mutation accumulation poses a threat to small asexual populations as purifying selection is too weak to remove the continual accumulation of mutations, driving the population to lower and lower mean relative fitness and increase in among-line variance over time (Kibota and Lynch, 1996). Among-line variance of fitness can be attributed to drift, as the individual lines possess identical ancestors, and are subjected to identical experimental conditions (Travisano et al.,

1995). Because there are relatively more deleterious mutations compared to beneficial mutations (Eyre-Walker and Keightley, 2007; Keightley and Eyre-Walker, 2010), the unidirectional ratchet-like accumulation of mutations may result in population extinction as a result of mutational meltdown (Gabriel et al., 1993).

MA experiments provide information on the rates of deleterious mutations and their distribution sizes that are relevant for many evolutionary hypotheses, and when coupled with whole-genome sequencing, they can also be used to accurately measure spontaneous mutation rates (Kondrashov and Kondrashov, 2010). A more recent evolution experiment also suggested that MA experimental lines could be used to distinguish beneficial and deleterious mutations (Tenaillon et al., 2016), as bottlenecks eliminate the variation needed for natural selection and all types of mutations accumulate at the rate at which they happen, regardless of their fitness cost (Barrick et al., 2009) . However, it can be difficult to relate mutational change to function, and we, therefore, sought to examine this by combining a newly developed method known as Delta-bitscore (DBS) for identifying functional perturbation with mutation accumulation (Wheeler et al., 2016). Tenaillon and colleagues (2016) predicted the impact of mutations by calculating the observed non-synonymous mutations relative to gene length, but this method does not take into account the effect of mutations on protein function. However, DBS analysis assesses the severity of observed mutations on protein function based on profile-HMM models built using orthologous protein sequences, allowing us to predict the functional divergence of the protein sequence.

Using fitness measurements and whole-genome sequencing, we examined the evolutionary dynamics of 10 independent mutation accumulation lines of a hypermutator *E. coli* population. For approximately 4,000 generations, 10 populations of hypermutator *E. coli* have been

subjected to 100 single-colony bottlenecks, and have been allowed to adapt to a nutrient rich solid media. We performed whole-genome sequencing to investigate the significance of genetic drift in driving the phenotypic and genotypic evolution of *E. coli*. It was anticipated that Muller's ratchet would be operating efficiently under these conditions, which would result in the accumulation of mutations and a relative loss of fitness. We used the DBS method to identify putative losses of protein function in these *E. coli* populations allowing us to assess the severity of the mutations on protein function and predict whether mutations with large DBS may have contributed to the loss of cell fitness. This combination of experimental evolution and whole-genome sequencing enables us to study the evolutionary trajectory and dynamics of *E. coli* populations under conditions of drift.

Methods

Strains and media

All chemicals were purchased from Sigma-Aldrich Co. unless otherwise specified. All oligonucleotides were synthesised by Integrated DNA Technologies. *E. coli* B strain REL606 was obtained from T. Cooper (University of Houston, Texas). REL606 (genotype: F^-, tsx-467(Am), araA230, lon^-, rpsL227(strR), $hsdR^-$, [mal^+](LamS) and REL606-derived strains were grown at 37°C in Luria Bertani (LB) media (Oxoid). For solid media, bacteriological agar (Oxoid) was added to a final concentration of 1.5% w/v. All the experiments were conducted in the presence of antibiotics at the following concentrations: streptomycin, 100µg/mL and ampicillin, 100µg/mL (Peptides International).

Introduction of a dominant mutator allele

A pGEM-T Easy (Promega) plasmid bearing a dominant *mutD5* mutation in the *dnaQ* gene, which encodes the 3'-5' proofreading exonuclease of DNA polymerase III holoenzyme, was introduced into REL606 by transformation (Sambrook et al., 1989). The spontaneous mutation rate of *E. coli* is approximately 1×10^{-3} per genome per replication, and this rate may increase 10^4 to 10^5 fold in the presence of the *mutD5* (Cox, 1976; Degnen and Cox, 1974).

Mutation accumulation experiment

A total of 10 genetically identical lineages were derived from a single glycerol stock of REL606 that contained a pGEM::*mutD5* plasmid. Single colonies of each individual lineage were randomly picked, excised from the agar, re-suspended in 15% glycerol and streaked onto fresh Luria Bertani (LB) agar and grown at 37˚C for 24 hours. The pick-streak-incubate process (growth cycle) was repeated for 100 growth cycles, where the cells were maintained on LB agar plates supplemented with streptomycin and ampicillin. To ensure random selection of

colonies, the last single colony of the streak, regardless of size, was selected. Five non-bottleneck control lines were also propagated through daily streaking of 100 uL of washed cells (agar was washed with 1 mL of 1X PBS) onto fresh LB agar (Figure 2.1). A glycerol stock of each MA lineage was prepared every day (bottlenecked). Glycerol stocks of every tenth passage of control non-bottlenecked lineages were also prepared.

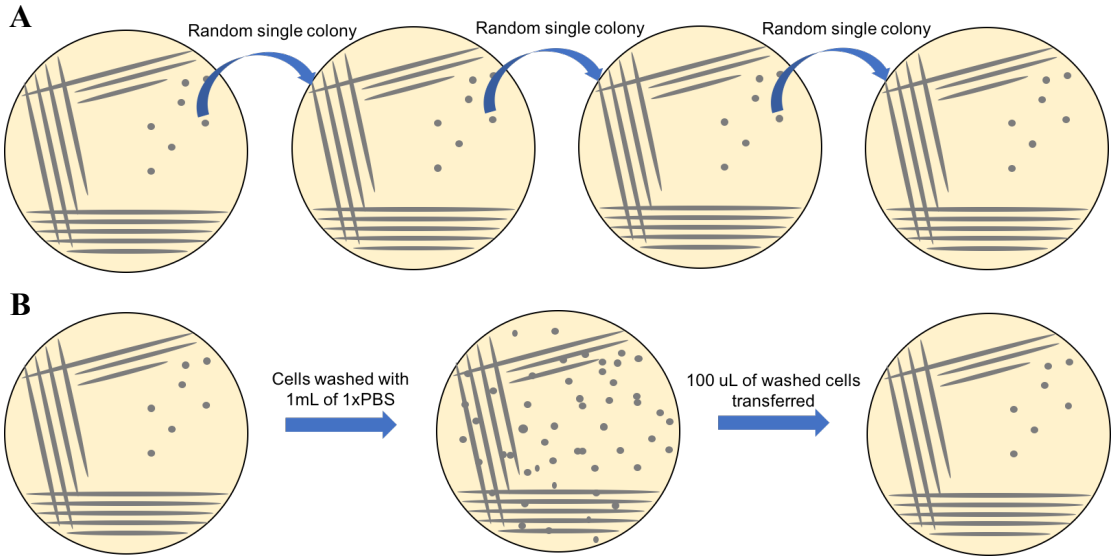

Figure 2.1 | Experimental design of the MA experiment. A) The MA lines were put through single-colony bottlenecks, where a random single colony was picked and streaked onto a fresh LB agar plate. **B)** For the non-bottlenecked control lines, the plates were washed with 1 mL of 1xPBS and 100 uL of washed cells were streaked onto a fresh LB agar plate.

Determination of growth rate

Cultures of all of the lineages of REL606 + pGEM::*mutD5* subjected to the MA experiment, along with the control lineages, were cultured every tenth day until the cultures reached saturation in LB containing streptomycin and ampicillin. The saturated cultures were then diluted 1:100 in 2 mL of fresh LB to an OD_{595} of 0.03-0.04, and distributed into a 24-well cell culture plate. Each experimental run consisted of 3 technical replicates per lineage and a

negative control (fresh LB). The OD_{595} of these cultures were then monitored for a period of 24-48 hours at 37˚C (with shaking at 200 rpm), taking OD_{595} measurements every 6 minutes using a FLUOstar Omega Microplate Reader (BMG Labtech). The growth rates for each line (averaged across replicate cultures) were determined as the minimum doubling time taken over a 30-minute interval.

Whole genome sequencing

At the end of 100 transfers, the bottlenecked and non-bottlenecked lineages were streaked to single colonies on LB agar. Single colonies were then used to inoculate LB liquid media. Genomic DNA was isolated using the Wizard Genomic DNA Purification Kit (Promega) and quantified using the Nanodrop 1000 Spectrophotometer and Qubit 2.0 Fluorometer. Sequencing was carried out by Macrogen (Korea) using an Illumina MiSeq platform with 2x250bp paired-end reads. Raw sequencing data were analysed using an in-house pipeline. Briefly, the sequencing reads were processed using AdapterRemoval (Lindgreen, 2012) to remove low-quality reads and adapter sequences. Reads were then mapped to the REL606 genome (NC_012967) using Bowtie2 (Langmead and Salzberg, 2012) (using default parameters, specifying haploid genomes where necessary), and the mapped genomes were visualized in Geneious v9.1.3 (Kearse et al., 2012). Genotyping was then carried out using SNPest (Lindgreen et al., 2014).

Delta-bitscore (DBS) analysis

The sequenced, mapped and genotyped genomes were translated to whole proteome sequences in Geneious v9.1.3 (Kearse et al., 2012) and these protein sequences were used for the delta-bitscore (DBS) calculations. DBS analyses utilise a profile Hidden Markov Model (HMM)-based approach that captures information on the expected frequency of occurrence of different

amino acids, insertions, and deletions across an alignment of protein sequences (Wheeler et al., 2016). HMM profile models for gamma-proteobacterial protein sequences were retrieved from the EggNOG database (Huerta-Cepas et al., 2016). Each of our protein sequences was aligned to their respective profile HMM to produce bitscore values, and by subtracting the bitscores of ancestor protein (reference) from that of the evolved protein (variant), a measure of divergence between the proteins is produced, termed DBS (Wheeler et al., 2016). DBS is an indication of how well the evolved protein sequence fits the sequence constraints modelled by the HMM, relative to the ancestral sequence. Mutations observed in highly conserved regions are likely to be less tolerated and are scored most harshly than those in non-conserved regions, and would thus give us an indication of the severity of the mutations (based on calculated DBS) we are seeing, and whether these are likely to be impacting protein function. A DBS value of 5 and above is considered to be deleterious to protein function.

Results

Genome sequencing reveals the accumulation of mutations

To create conditions favouring drift, we subjected ten independent *E. coli* REL606 + pGEM::*mutD5* lines to a 100 single-colony passages. After approximately 4,000 generations, the genomes of these ten independent bottlenecked lineages (BN100.1-BN100.10), and five control lineages (C100.1-C100.5) that were not subjected to genetic bottlenecks, were sequenced. All 15 lineages were multiplexed in a single Illumina MiSeq Lane, resulting in 2.1 Gb of data across 9.3 million reads. The sequence data was of high quality, with 87% of bases having a Q30 value or higher. We also sequenced the REL606 + pGEM::*mutD5* ancestor line (Day 0) to produce a reference genome. The reads of the evolved lineages were then mapped to the reference genome. The mapping gave a sequencing depth of 32±20x ($\bar{X}$±SD) across all the lineages.

The bottlenecked lineages excluding BN100.1 (which accumulated approximately 13-fold more mutations compared to the other bottlenecked lineages) accumulated an average of 351±124 ($\bar{X}$±SD) mutations, while the non-bottlenecked lineages accumulated an average of 232±48 ($\bar{X}$±SD) mutations (Table 2.1). On average, the bottlenecked lineages accumulated more mutations than the non-bottlenecked lineages (P-value = 0.019, Wilcoxon rank sum).

Table 2.1 | The number of SNPs in the bottlenecked and non-bottlenecked lineages after 100 passages

Lineage	Number of substitutions	Number of insertions	Number of deletions	Total number of mutations
BN100.1	4433	211	159	4803
BN100.2	371	5	9	385
BN100.3	287	10	6	303
BN100.4	216	6	7	229
BN100.5	339	7	2	348
BN100.6	578	33	35	646
BN100.7	234	6	7	247
BN100.8	323	9	7	339
BN100.9	345	18	19	382
BN100.10	268	6	4	278
C100.1	283	10	6	299
C100.2	224	7	4	235
C100.3	214	3	3	220
C100.4	229	11	14	254
C100.5	141	3	7	151

BN represents the bottlenecked lineages while C represents the non-bottlenecked (control) lineages. The numbers following BN and C denote the number of passages followed by the randomly assigned lineage number respectively.

Mutational bias was observed for substitution mutations

A bias towards substitution mutations was observed in every lineage, with substitution mutations constituting approximately 90% of the total mutations (Table 2.1). The frequency of substitution mutations (transitions and transversions) was relatively higher than indels (insertions and deletion) in both the bottlenecked and non-bottlenecked lineages (Figure 2.2).

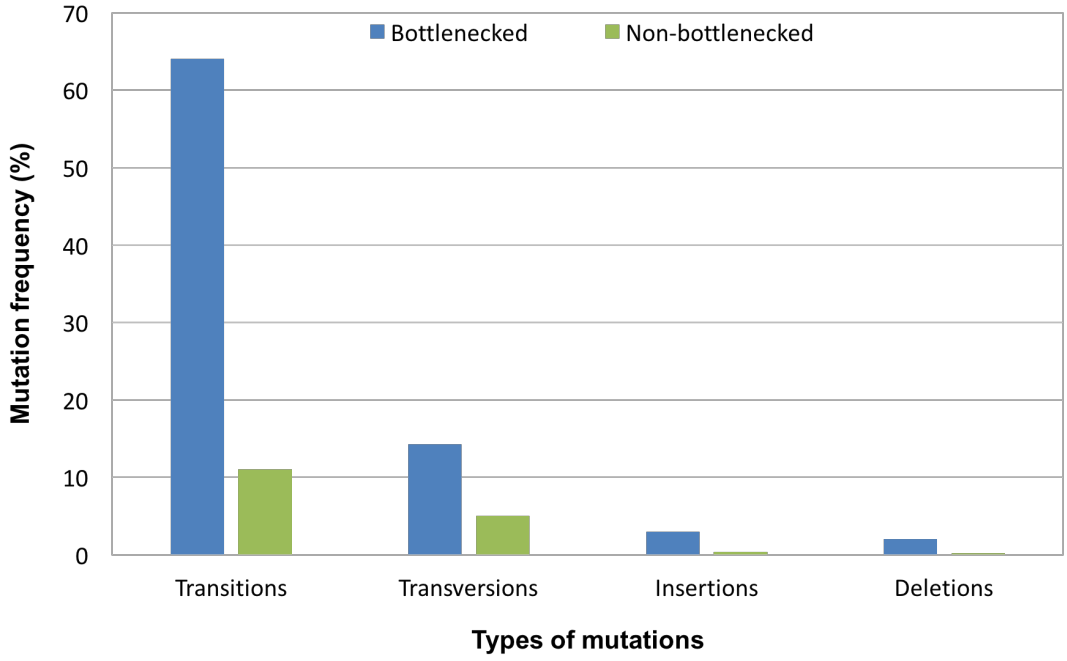

Figure 2.2 | The frequency and types of mutations observed in the bottlenecked and non-bottlenecked lineages. The bottlenecked lineages showed a consistently higher frequency of mutations over all types of mutations observed compared to the non-bottlenecked lineages. Transitions make up approximately 75% of the total mutations while indels were rarely observed relative to substitution mutations.

The emergence of additional hypermutator alleles

To increase the mutation rate, a *mutD5* mutator allele was first introduced into the ancestor lineage. By the end of the experiment, a decrease in mutation rate was observed for all the lineages of the MA experiment compared to the ancestor (Table 2.2). However, these calculated mutation rates are higher than the documented wild-type *E. coli* rate of 10^{-10} per nucleotide per generation (Foster et al., 2015; Lee et al., 2012). On average, the mutation rate of the bottlenecked lineages is higher than the non-bottlenecked lineages (*P*-value = 0.012, Wilcoxon rank-sum). Upon whole-genome mapping, it was observed that the introduced

mutD5 mutator allele remained in all the lineages. In addition, mutations in the *mutS*, *mutT* and *mutY* genes were observed in one of the 10 bottlenecked lineages (BN100.1).

Table 2.2 The calculated mutation rates of the MA lineages

Lineage	Number of generations	Mutation rate per nucleotide per generation
Ancestor	44	7.27×10^{-7}
BN100.1	<u>2666</u>	3.89×10^{-7}
BN100.2	3837	2.16×10^{-8}
BN100.3	4152	1.58×10^{-8}
BN100.4	4348	1.14×10^{-8}
BN100.5	3890	1.92×10^{-8}
BN100.6	4366	3.20×10^{-8}
BN100.7	4015	1.33×10^{-8}
BN100.8	4547	1.61×10^{-8}
BN100.9	3344	2.47×10^{-8}
BN100.10	5016	1.20×10^{-8}
C100.1	4556	1.42×10^{-8}
C100.2	4648	1.09×10^{-8}
C100.3	4635	1.03×10^{-8}
C100.4	4171	1.32×10^{-8}
C100.5	4475	7.29×10^{-9}

The number of generations was calculated based on the doubling time calculated for every transfer (24 hours) for the duration of the evolution experiment. The mutation rate was calculated by dividing the number of mutations by the number of generations and number of nucleotides in the genome of *E. coli* REL606. Number underlined represents the calculated number of generations based on the last measurable growth rate at day 50 (BN50.1).

One bottlenecked lineage succumbed to mutational meltdown

In addition to monitoring growth rates, we also visually inspected colony size every day for all of the lineages. We observed a reduction in colony size from day 10 onwards of BN100.1 (Figure 2.3). This phenotypic change continued throughout the experiment and did not revert back to the ancestral state. The measured growth rates were consistent with the observed reduction in colony size (results show day 20 onwards), where the doubling time increased from 30.6±0.6 min to 41.2 min, 81.5±3.3 and 82.0±3.4 min (n=3, $\bar{X}$±SE) from day 10 to day 40. This bottlenecked lineage was not revivable from glycerol stocks after 50 days.

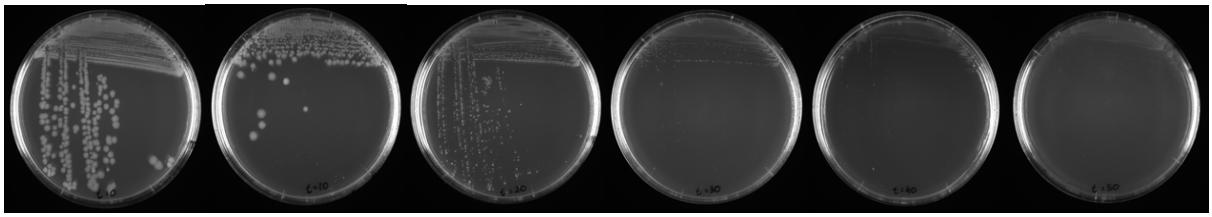

Figure 2.3 | Colony size decreased over time in one of the bottlenecked lineage, BN100.1. From left to right, each plate depicts the serial single-colony bottleneck in 10-days intervals from day zero to day 50. The average colony size decreased from day 20 onwards until the end of the experiment without recovering back to the ancestral colony size.

An overall loss of fitness was observed in all the lineages

The growth rates of the evolved lineages were measured at 10-day intervals using a plate reader. A decrease in average growth rate (calculated from doubling time) relative to the ancestor was observed for both the bottlenecked which and non-bottlenecked lineages (Figure 2.4). It is important to note that growth rate calculations for BN100.1 were excluded from the calculation of average growth rate, as these lines succumbed to mutational meltdown.

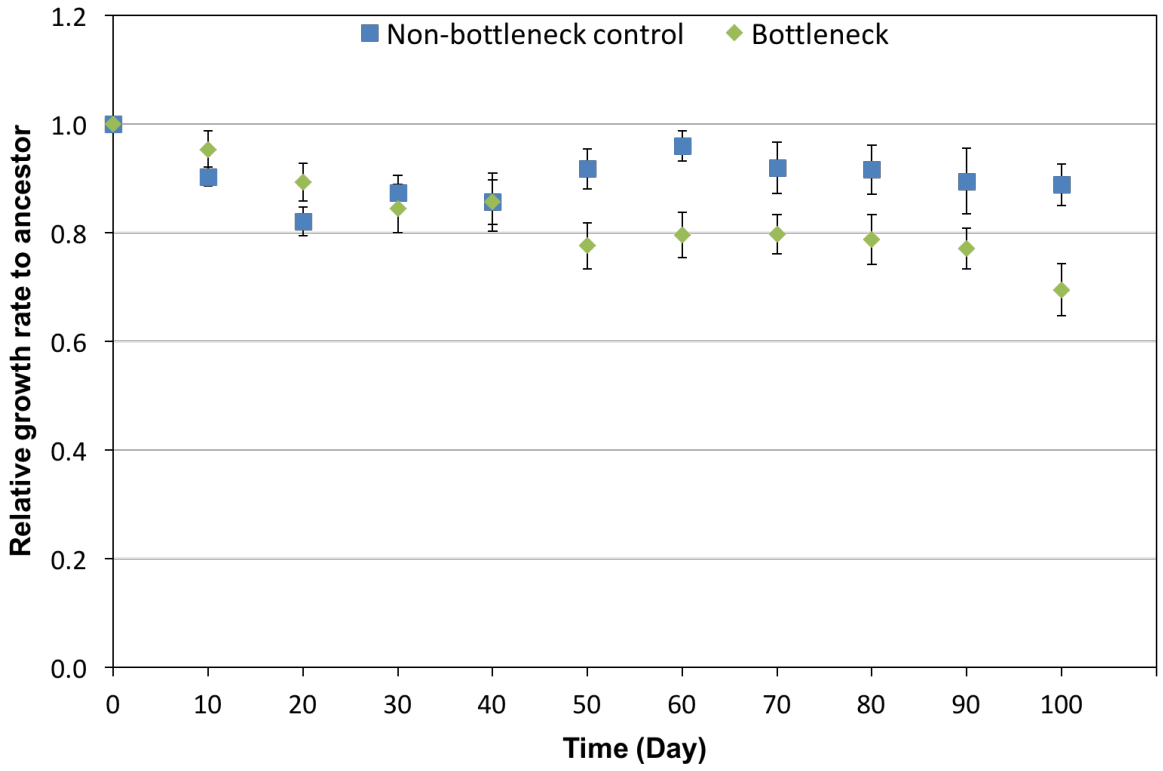

Figure 2.4 | The average relative growth rates of the non-bottlenecked and bottlenecked lineages decreased with time compared to the ancestor. Each point represents the relative growth rate compared to the REL606 + pGEM::*mutD5* ancestor. The average relative growth rate of the non-bottlenecked controls (blue squares) (n=5, $\bar{X}$±SE) and bottlenecked lines (green diamonds) (n=9, $\bar{X}$±SE) decreased with time. Values above 1 represent an increase in growth rate while values below 1 represent a decrease in growth rate compared to the ancestor.

During the course of our MA experiment, the bottlenecked lineages suffered a loss of fitness (measured as doubling time) relative to the ancestor (Figure 2.5). By the end of the evolution experiment, the doubling time of the bottlenecked lineages increased to 35.4-53.7 min from 32.7 min calculated for the ancestor. On average, cell fitness of the bottlenecked lineages decreased with time, with the exception of BN100.10 where we observed an apparent fitness gain compared to the ancestor. We also note that BN100.1 was excluded from this analysis as it was not revivable from its frozen glycerol stock.

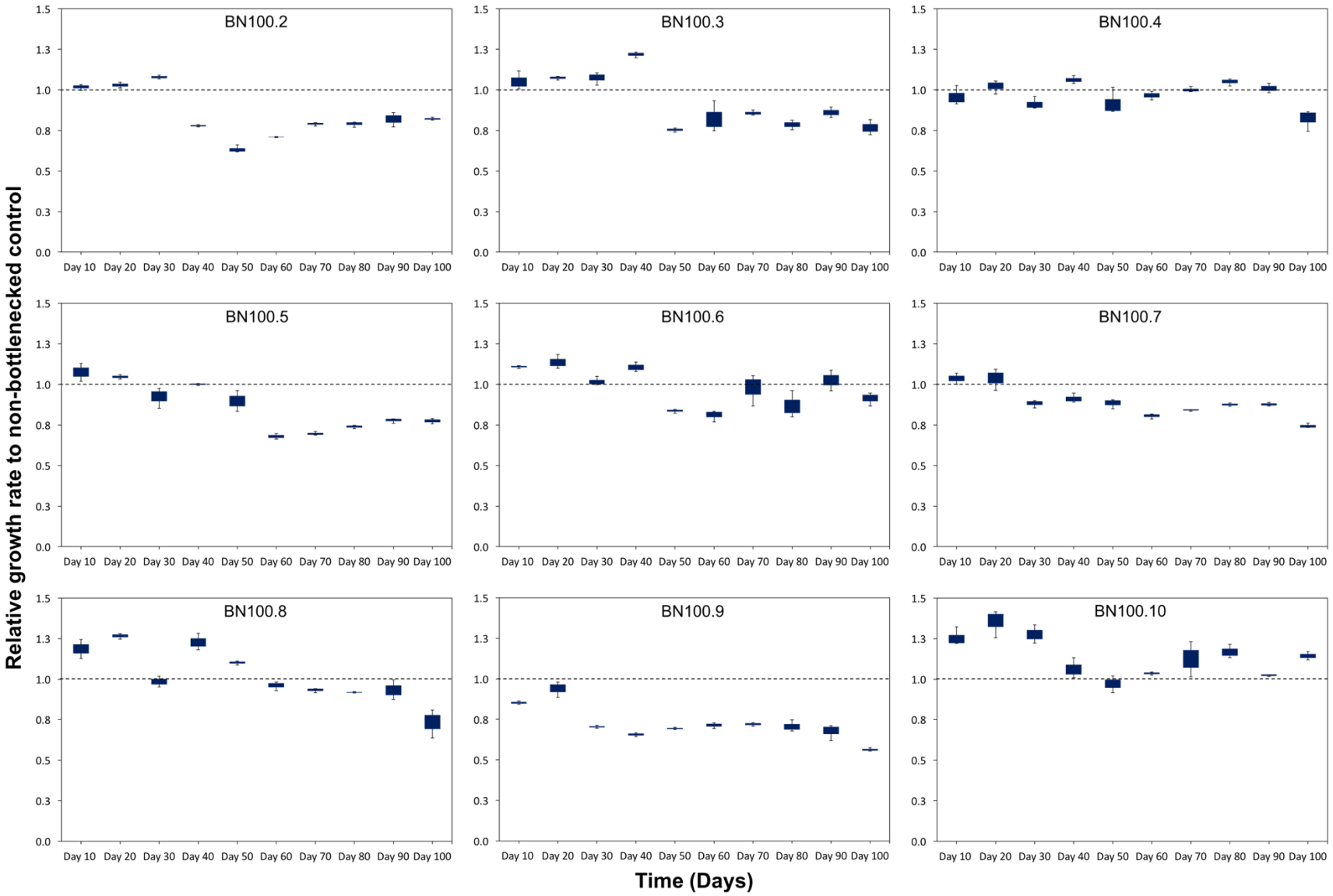
BN100.2
BN100.3
BN100.4
BN100.5
BN100.6
BN100.7
BN100.8
BN100.9
BN100.10
Relative growth rate to non-bottlenecked control
Time (Days)
1.5
1.3
1.0
0.8
0.5
0.3
0.0
Day 10 Day 20 Day 30 Day 40 Day 50 Day 60 Day 70 Day 80 Day 90 Day 100

Figure 2.5 | The average relative growth rates of the bottlenecked lineages decreased compared to the non-bottlenecked control. Each point represents the relative growth rate at ten-day intervals compared to the control. The growth rate was measured as the minimum doubling time over a 30-minute interval. The average growth rate is shown as relative to that of the average growth rate of the control lineages (n=3, $\bar{X}$±SE). Values above 1 represent an increase in growth rate while values below 1 represent a decrease in growth rate relative to the control.

We also measured the growth rates of the non-bottlenecked control lineages and observed an average loss of fitness (measured as doubling time) compared to the ancestor (Figure 2.6). The doubling time of the non-bottlenecked lineages on average increased to 32.7-39.9 min from 32.7 min calculated for the ancestor by the end of the evolution experiment.

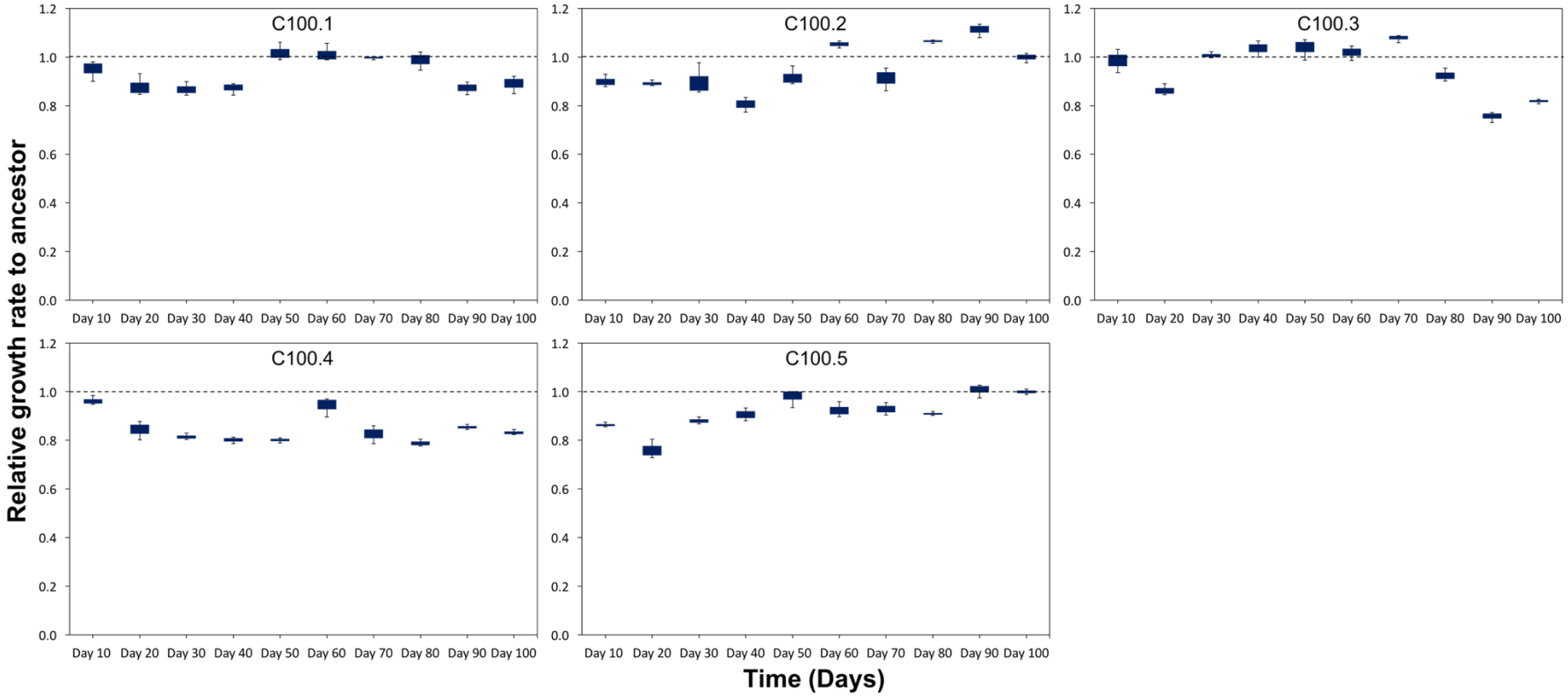

Figure 2.6 | The relative growth rates of the non-bottlenecked lineages decreased compared to the ancestor. Each point represents the relative growth rate at ten-day intervals compared to the REL606 + pGEM::*mutD5* ancestor. The growth rate was measured as the minimum doubling time over a 30-minute interval. The average growth rate is shown as relative to that of the average growth rate of the control lineages (n=3, $\bar{X}$±SE). Values above 1 represent an increase in growth rate while values below 1 represent a decrease in growth rate relative to the control.

We then compared the growth rates of the bottlenecked to non-bottlenecked lineages to rule out fitness changes being attributed to media adaptation. On average, the bottlenecked lineages were less fit (measured as doubling time) compared to the non-bottlenecked lineages by the end of the MA experiment, with the exception of BN100.10 (Figure 2.7). The average doubling time of the bottlenecked lineages increased to 46.4±1.9 from 36.3±0.8 min (n=9, $\bar{X}$±SE) calculated for the non-bottlenecked control lineages (*P*-value = 4.05 x 10^{-5}, Wilcoxon rank sum).

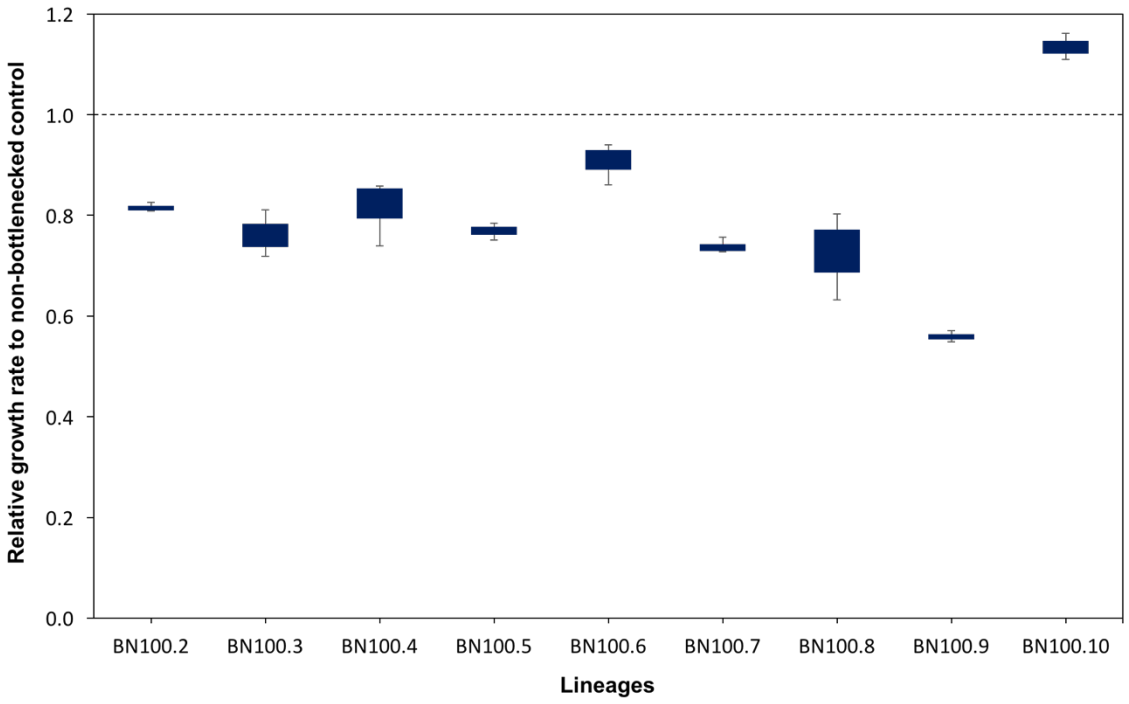

Figure 2.7 | The growth rates of the bottlenecked lineages decreased compared to the non-bottlenecked lineages in the MA experiment. Each point represents the relative growth rate of the bottlenecked lineages compared to the non-bottlenecked control lineages at the end of the evolution experiment. The growth rate was measured as the minimum doubling time over a 30-minute interval. The average growth rate is shown as relative to that of the average growth rate of the control lineages (n=3, $\bar{X}$±SE). Values above 1 represent an increase in growth rate while values below 1 represent a decrease in growth rate relative to the control.

DBS analysis predicted loss of protein function

Given the observed loss of fitness in the bottlenecked lineages upon growth rate measurements and visual colony size inspection, we investigated whether the accumulated mutations could have contributed to the loss of fitness associated with loss of protein function by calculating the DBS. DBS values were calculated at the end of the evolution experiment by subtracting the bitscore of the evolved from the ancestral protein sequences. We observed a large value of the sum of absolute DBS (∑|DBS|) in BN100.1 compared to the other bottlenecked lineages (Table 2.3). When we compared the ∑|DBS| of the bottlenecked (excluding BN100.1) to the non-bottlenecked control lineages, we observed no significant difference (*P*-value = 0.254, Wilcoxon rank-sum).

Table 2.3 | The ∑|DBS| and doubling time at the end of the evolution experiment

Lineage	∑|DBS|	Doubling time (minutes) (n=3, $\overline{X}$±SE)
BN100.1	29320	*
BN100.2	626	44.1±0.3
BN100.3	1245	47.4±1.7
BN100.4	591	44.3±2.2
BN100.5	500	46.9±0.6
BN100.6	3045	39.8±1.1
BN100.7	1295	48.8±0.6
BN100.8	1616	50.2±3.6
BN100.9	1843	64.4±0.7
BN100.10	1071	31.7±0.4
C100.1	1395	36.8±0.9
C100.2	1268	32.7±0.4
C100.3	364	40.0±0.3
C100.4	975	39.3±0.3
C100.5	560	32.7±0.2

*No growth data were collected for BN100.1 as it was not revivable from glycerol stock

Sum of absolute DBS does not correlate to cell fitness

To examine whether a loss of protein function could have contributed to the loss of fitness observed, we performed a Pearson's correlation test. Although a large ∑|DBS| was observed in the majority of the bottlenecked lineages, the ∑|DBS| does not correlate to the loss of fitness measured as doubling time (R^2 = 0.0089, *P*-value = 0.8092) (Figure 2.8). This was also observed in the non-bottlenecked control lineages where the calculate ∑|DBS| does not correlate to doubling time (R^2 = 0.0664, *P*-value = 0.6755) (Figure 2.8).

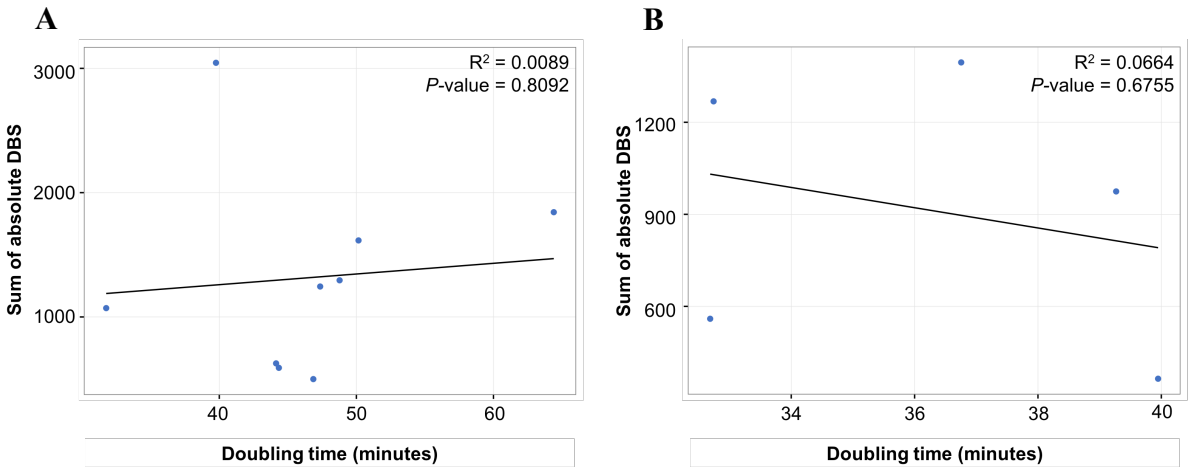

Figure 2.8 | Scatter plot of the correlation between and sum of absolute DBS and doubling time. The x-axis represents the doubling time and the y-axis represents the sum of DBS. **A)** The ∑|DBS| and doubling time of the bottlenecked lineages (excluding BN100.1 which succumbed to mutational meltdown) were calculated after 100 serial passages. **B)** The ∑|DBS| and doubling time of the non-bottlenecked lineages were calculated after 100 passages. The calculated ∑|DBS| does not correlate to the doubling time for both the bottleneck and non-bottlenecked lineages.

As we observed no correlation between the ∑|DBS| and doubling time, we plotted individual DBS values of each lineage to examine whether individual mutations may have contributed to the observed decline in cell fitness (Figure 2.9). Although the majority of mutations acquired

in each lineage have low DBS values, we observed some outliers with high DBS for lineages BN100.1, BN100.7, BN100.9, and C100.2. The genes which possess extreme DBS values were *yicI*, *yghV*, *yrfF* (B100.1), trkG (BN100.7), *yrfF* (BN100.9) and *ycbS* (C100.2).

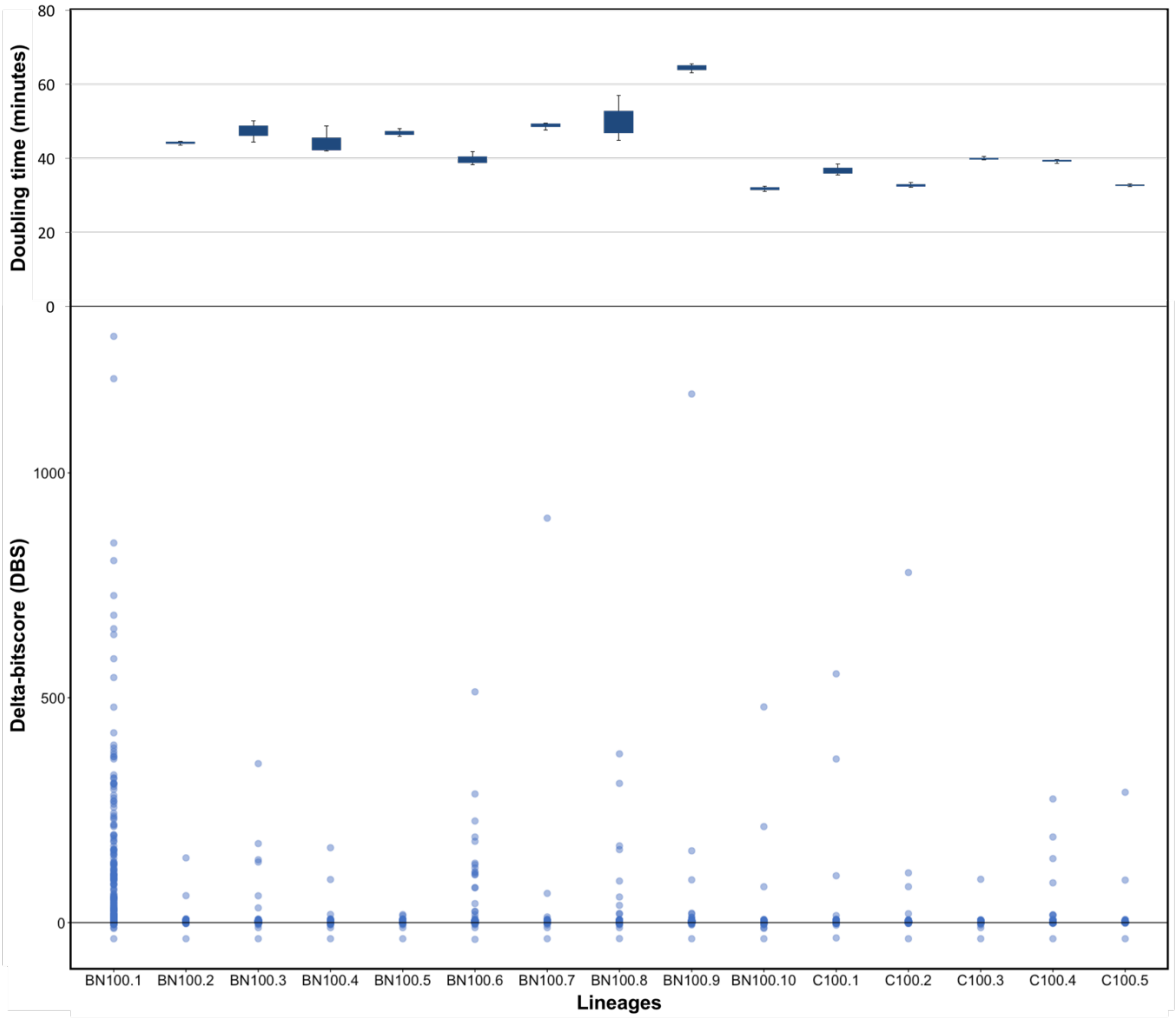

Figure 2.9 | Doubling time and DBS of the bottlenecked lineages. The top graph indicates the doubling time of the lineages at day 100. No data on growth rate was collected for BN100.1 as it succumbed to extinction. The bottom graph represents DBS calculated where every blue circle represents an individual DBS for each protein sequence. DBS was calculated by subtracting the bitscore of the evolved from that of the ancestor bitscore. Multiple outliers (high DBS) were observed for BN100.1 while an outlier each was observed for BN100.7, BN100.9, and C100.2.

Discussion

We subjected ten independent *E. coli* populations to single-colony bottlenecks and sequenced the whole genomes of the 10 bottlenecked lineages, and 5 non-bottlenecked control lineages following up to 4,000 generations of culturing. Previous MA experiments have focused mainly on the distribution of mutations, mutation rate, and the effect of mutations on fitness but to our knowledge, there have been no studies that directly examine the effect of mutational changes on protein function. Here, we examined the effect of mutations on protein function and predicted the severity of mutations on protein function using DBS.

Following approximately 4,000 generations of culturing, we observed an increase in the number of mutations for all the lineages, including the non-bottlenecked control lineages, when compared to the ancestor (Table 2.1 and Figure 2.2). The number of mutations in the bottlenecked lineages was significantly higher than the control lineages (*P*-value = 0.012, Wilcoxon rank-sum), consistent with previous MA experiments (Andersson and Hughes, 1996; Tenaillon et al., 2016). Muller's ratchet resulted in an unidirectional accumulation of slightly deleterious mutations in the absence of genetic recombination and selection in populations subjected to bottlenecks (Felsenstein, 1974; Muller, 1964). The difference between the number of mutations accumulated in the bottlenecked and non-bottlenecked lineages may be accounted for by the more efficient negative selection in the non-bottlenecked lineages (Nielsen, 2005). Furthermore, few parallel mutations were observed across all the lineages, suggesting that purifying selection is weak and the mutational trajectory is non-adaptive, as parallel evolution is a hallmark of adaptation (Wielgoss et al., 2013).

We then analysed the nature of the mutations and observed a bias in the mutational spectrum towards A.T base pairs. A study showed that sequence evolution of small populations is

affected by mutational biases, as the efficacy of selection and biased gene conversion (BGC) are severely reduced relative to drift (Nagylaki, 1983). This may explain the mutational bias towards A.T that we observed in our MA experiment, as also observed in other clonal pathogens that were subjected to genetic bottlenecks (Hershberg and Petrov, 2010; Lind and Andersson, 2008). The conditions of our experimental evolution were designed to mimic the genetic structure of *Buchnera* with high drift and low selection (see Chapter 1). A bias towards an AT rich genome has also been observed in *Buchnera* (Shigenobu et al., 2000) a close relative of *E. coli* (Baumann et al., 2006). Although, the genomic AT content of our bottlenecked lineages remained at approximately 49.2%, (as observed in the ancestor), we observed a mutational bias towards A and T substitutions.

Next, by measuring the mutation rate, we asked whether mutation may dictate the tempo and mode of evolution under efficient drift. The calculated average mutation rate of the bottlenecked lineages was 5.5×10^{-8} per nucleotide, per generation, and this average rate of spontaneous mutation is 5-fold higher than the control lineages (Table 2.2). Although the calculated mutation rates were higher than wild-type *E. coli* at 10^{-10} (Foster et al., 2015; Lee et al., 2012), these observed mutation rates were lower than the hypermutator ancestor. The decrease in mutation rate could be explained by the tension between adaptation and genetic load of hypermutator strains. It has been documented that a higher mutation rate may be favoured when the populations are subjected to a new environment as they provide more beneficial mutations allowing a more rapid response to selection (André and Godelle, 2006; Taddei et al., 1997). However, as the population becomes well-adapted to the current environment, the mutation rate would be lowered to relax the genetic load (Kimura, 1967). A study by Wielgoss et al. (2013) demonstrated that reduction in the mutation rate of hypermutators was directed by selective advantage, as the potential for further adaptation

declined with time. The reduction in mutation rate observed in our MA experiment could be explained by this selective advantage (Table 2.2). Even though our experimental conditions were specifically designed to favour drift and minimise selection, selection can not be completely eliminated.

Subsequently, we determined the relative fitness of the serially single-colony passaged lineages by measuring the growth rates. The average relative mean fitness of both the bottlenecked and non-bottlenecked lineages progressively decreased relative to the ancestor (Figure 2.4). Further to this, the average relative growth rates of the bottlenecked lineages is also lower compared to the non-bottlenecked control (Figure 2.5), and by the end of the experiment, 8 of the 9 bottlenecked lineages grew slower compared to the non-bottlenecked lineages (*P*-value = 4.05 x 10^{-5}, Wilcoxon rank sum) (Figure 2.7). The calculated doubling time of the bottlenecked lineages ranged from 39.8 to 64.4 minutes, compared to 36.3 minutes for the non-bottlenecked lineages and 32.7 minutes for the ancestor. These results are consistent with previous studies, where the accumulation of deleterious mutations coupled with a defective DNA mismatch repair (MMR) system resulted in a loss of fitness (Funchain et al., 2000; Sniegowski et al., 1997). All of the lineages (bottlenecked and non-bottlenecked) retained the mutated *mutD5* allele that results in a defective MMR system. On average, a net loss of fitness was also observed in the non-bottlenecked lineages. While a defective MMR system does not necessarily result in a loss of fitness, when combined with bottlenecking conditions, any deleterious mutations that do occur are more likely to be fixed in the population by chance as observed in the bottlenecked lineages.

BN100.1 suffered from a major fitness decline as observed by a decrease in colony size (Figure 2.3) and extinction recurred consistently in replicate populations re-established from the

glycerol stock but not in populations initiated from the pre-decline stock. This particular lineage had accumulated approximately 13-fold more mutations than the average number of mutations accumulated in the other 9 bottlenecked lineages (Table 2.1), and the growth rate could not be determined after the 40th single-colony passage. As there are more deleterious mutations than beneficial mutations (Eyre-Walker and Keightley, 2007), and the advantage gained by lower mutation rate cannot compensate for the effect of genetic drift in small asexual populations (Lynch, 2010), we would expect to see a fitness decline with the accumulation of mutations. The documented fitness decline is concurrent with previous MA studies in bacteria (Andersson and Hughes, 1996), endosymbiont (Moran, 1996), yeast (Zeyl et al., 2001) and virus (Chao, 1990; Duarte et al., 1992). This positive feedback of accumulation of mutations and fitness decline which resulted in mutational meltdown of BN100.1 (cells were not revivable after Day 50) has also been observed in laboratory populations of yeast (Zeyl et al., 2001), providing evidence that under conditions of high drift, small populations may eventually become extinct as a result of increasing mutational load (Gabriel and Bürger, 1994; Lynch and Gabriel, 1990; Lynch et al., 1995).

One of the bottlenecked lineages, BN100.10 was fitter than the ancestor and control by the end of the evolution experiment. Although the number of deleterious mutations outnumbered the number of beneficial mutations (Eyre-Walker and Keightley, 2007; Keightley and Eyre-Walker, 2010), the gain in fitness observed in this lineage could be attributed to either a small number of compensatory or beneficial mutations in this lineage, or that the accumulated mutations were not located in essential genes. Although most deleterious mutations are negatively selected, occasionally some may revert back to wild-type or become compensated by mutation elsewhere in the genome, thus recovering fitness (Szamecz et al., 2014).

We observed an increase in the number of mutations and an average decrease in fitness in our MA lineages. To test whether the decrease in fitness could be attributed to a loss of protein function, we assessed the severity of the observed mutations on protein function using DBS. We observed an increase in the sum of absolute delta-bitscore, ∑|DBS| in our MA lineages compared to the ancestor (Table 2.3). However, there is no significant difference between the ∑|DBS| of the bottlenecked and non-bottlenecked control lineages (*P*-value = 0.254, Wilcoxon rank-sum), although on average, the bottlenecked lineages were less fit compared to the non-bottlenecked control lineages. This result suggests that the ∑|DBS| does not directly correlate to the observed fitness loss (measured as doubling time) (Figure 2.8, R^2 = 0.044, *P*-value = 0.47). Nevertheless, we observed a notably higher number of mutations and ∑|DBS| in one of the 10 bottlenecked lineages (BN100.1) compared to the other bottlenecked lineages. This may be explained by the stochasticity in the types of mutations that were accumulated leading to a greater divergence in the number of mutation accumulated among lineages (Kibota and Lynch, 1996).

As aforementioned, BN100.1 succumbed to mutational meltdown, and the high ∑|DBS| may explain this outcome as most of the mutations accumulated may have deleterious effects on protein function. This particular lineage had a calculated ∑|DBS| of approximately 30,000 which is 22-fold larger than the average ∑|DBS| of the other bottlenecked lineages (Table 2.3). One of the frameshift mutations located within the *rrmA* gene (DBS of 394.7, *P*-value = 0) that codes for the rRNA large subunit methyltransferase responsible for the methylation of the 23S rRNA (Gustafsson and Persson, 1998). Interestingly, the mutation observed was within a polyT tract where an addition of a T resulted in a frameshifted poly 8T tract. Studies have shown that long polyA/T are prone to slippage-type editing which would result in a heterogeneous population of mRNAs with varying reading frames and lengths (Tamas et al., 2008; Wagner et

al., 1990). Slippage-type editing will ultimately result in a proportion of functional and non-functional proteins which may affect cell fitness under certain circumstances (discussed in Chapter 3 and 4). *rrmA* mutants have been shown to exhibit defects in translation, decreased growth rate and increased resistance to the ribosome binding antibiotic viomycin (Gustafsson and Persson, 1998). These findings suggest that the observed frameshift mutation in *rrmA* may be partly responsible for the loss of fitness in BN100.1.

We also observed a frameshift mutation in the *mutT* gene in BN100.1 that would result in a truncated 8-oxo-dGTP diphosphatase protein (DBS of 96.1, *P*-value = 1.1 x 10^{-300}). Furthermore, we observed non-synonymous mutations to the *mutS* and *mutY* genes and although the calculated DBS for both the mutations were insignificant, changes to the amino acid may affect protein folding and protein-protein interactions (Yates and Sternberg, 2013) potentially altering protein function. MutT, MutS, and MutY play a role in DNA mismatch repair (Marinus, 2010) and bacteria with defective MMR systems suffer from elevated mutation rates and an increase in recombination (LeClerc et al., 1996; Li, 2008). MutS and MutY defects have been shown to increase mutation rates in *E. coli,* where fluctuation tests reported a moderate increase in mutation rate for *mutY* strains and a greater increase in mutation rate for *mutS* strains (Arjan et al., 1999). The inactivation of the MMR system in *E. coli* has been suggested to destabilise the genome as high mutation rate leads to multiple losses of function (Funchain et al., 2000). In addition, mutator lineages also tend to accumulate deleterious mutations at a higher rate than advantageous mutations (Kimura, 1967). Thus, the observed increase in the number of SNPs in BN100.1 (Table 2.1) and the higher mutation rate compared to the other bottlenecked lineages (Table 2.2) could be attributed to the additional *mutS, mutT,* and/or *mutY*-based mutator alleles.

Although the sum of absolute DBS observed for all the lineages (excluding BN100.1) does not directly correlate to cell fitness (Figure 2.8, $R^2 = 0.0089$, *P*-value = 0.8092), this result can be explained by the experimental conditions used. The genes with high DBS could be enriched in 'non-essential' genes which do not contribute towards fitness under our experimental conditions. For example, a high DBS was observed for AraC (DBS = 433, *P*-value = 0), but the mutation in this gene does not affect fitness in rich media (as used in our MA experiment), but a loss of fitness was observed when grown in minimal media supplemented with arabinose as the sole carbon source (discussed in Chapter 3). Furthermore, deleterious mutations observed at the genome level could be removed at the mRNA level via mRNA degradation (Rauhut and Klug, 1999; Schoenberg, 2007), or in some cases corrected at the RNA level via editing-type processes (discussed in Chapter 3 and 4) (Tamas et al., 2008; Wernegreen et al., 2010), therefore preventing the translation of truncated proteins. If the damaged mRNAs were not decayed or corrected and truncated proteins were subsequently produced, molecular chaperones and proteases may remove these proteins (Dougan et al., 2002; Mogk et al., 2011). The experimental conditions, location and types of mutations, and the life-span of mRNA and protein may, therefore, explain the lack of correlation between fitness and elevated ∑|DBS| in our evolution experiment.

Interestingly, BN100.9 with the longest doubling time (Figure 2.9) has a lower ∑|DBS| value of 1842 compared to BN100.6 (∑|DBS| = 3055) which has one of the shortest doubling times among the bottlenecked lineages (excluding BN100.10) (Table 2.3). This result reinforced the lack of correlation between ∑|DBS| and the observed loss of fitness. However, upon further inspection of the individual mutations in BN100.9, we observed some genes with high DBS values despite the low calculated ∑|DBS|, suggesting that the calculated ∑|DBS| does not account for the infrequent outliers that may have a large effect on protein function and cell

fitness. In BN100.9, a mutation to the *yrfF* gene resulted in a truncated protein (DBS = 1176, *P*-value = 0). The *E. coli yrfF* gene codes for a putative inner membrane protein with unknown function, however, it is categorized as an essential gene in *E. coli* in the PEC (Yamazaki et al., 2008) and KEIO (Baba et al., 2006) databases under certain conditions. This suggests that the d mutation in the *yrfF* gene may have contributed to the long doubling time observed in BN100.9 despite the small calculated ∑|DBS|. To test this, we could first knock-out the *yrfF* gene and determine whether the gene is essential for cell survival, and subsequently introduce the observed mutation into the ancestor strain to examine the effect of the mutation on cell fitness.

Although some of the loss of fitness that we observed may be attributed to the loss of protein function (as observed in BN100.1), the DBS values are not fully reflective of what we have seen experimentally. One of the limitations to the DBS analysis lies in the fact that DBS only takes into account the impact of mutations in protein coding sequences, thus neglecting the non-coding sequences. Although we observed numerous mutations in the untranslated regions (UTRs), non-coding regions and RNA genes in our evolution experiment lineages, the severity of mutations in these regions were not accounted for by DBS. It has been well documented that non-coding RNAs (ncRNA) play crucial roles in regulatory networks in bacterial stress responses (Altuvia et al., 1997; Sledjeski et al., 1996), quorum sensing (Lenz et al., 2004), plasmid and viral replication (Wagner and Simons, 1994), and in bacterial virulence (Sittka et al., 2007). Further to this, we observed mutations to multiple tRNA genes and the *rrfF* gene that codes for the 5S rRNA in BN100.1. Previous studies have shown that a decrease in concentration and availability of tRNAs influence cell fitness (Fedyunin et al., 2012; Wohlgemuth et al., 2013), while the deletion of the 5S rRNA gene resulted in a fitness loss

(Ammons et al., 1999). These observations could explain the lack of correlation between loss of fitness and ∑|DBS| observed in our evolution experiment (Figure 2.8).

To summarise, we observed the accumulation of mutations and a decrease in relative fitness in our MA experiment. Although fitness decline is concurrent with mutation accumulation, we observed a situation where a small number of mutations with large effect (large DBS value) may lead to loss of fitness, in contrast to the 'death by a thousand paper cuts' model. Further investigation with DBS may help shed light on this phenomenon. Further to this, we also observed the emergence of frameshifted polyA/T tracts, and this finding will be the focus of Chapter 3.

References

Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L., and Storz, G. (1997). A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. Cell *90*, 43–53.

Ammons, D., Rampersad, J., and Fox, G.E. (1999). 5S rRNA gene deletions cause an unexpectedly high fitness loss in *Escherichia coli*. Nucleic Acids Res. *27*, 637–642.

Andersson, D.I., and Hughes, D. (1996). Muller's ratchet decreases fitness of a DNA-based microbe. Proc. Natl. Acad. Sci. *93*, 906–907.

André, J.-B., and Godelle, B. (2006). The evolution of mutation rate in finite asexual populations. Genetics *172*, 611–626.

Arjan, G.J., de Visser, M., Zeyl, C.W., Gerrish, P.J., Blanchard, J.L., and Lenski, R.E. (1999). Diminishing returns from mutation supply rate in asexual populations. Science *283*, 404.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of *Escherichia col*i K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Barrick, J.E., Yu, D.S., Yoon, S.H., Jeong, H., Oh, T.K., Schneider, D., Lenski, R.E., and Kim, J.F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature *461*, 1243–1247.

Baumann, P., Moran, N.A., and Baumann, L. (2006). Bacteriocyte-associated endosymbionts of insects. In The Prokaryotes, M.D.P. Dr, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt, eds. (Springer New York), pp. 403–438.

Bell, G. (1988). Sex and Death in Protozoa: The History of Obsession (Cambridge University Press).

Cervera, H., and Elena, S.F. (2016). Genetic variation in fitness within a clonal population of a plant RNA virus. Virus Evol. *2*, vew006.

Chao, L. (1990). Fitness of RNA Virus Decreased by Muller's Ratchet. Nature *348*, 454–455.

Charlesworth, B., and Charlesworth, D. (1997). Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet Res (Camb). *70,* 63-73.

Cox, E.C. (1976). Bacterial mutator genes and the control of spontaneous mutation. Annu. Rev. Genet. *10*, 135–156.

Degnen, G.E., and Cox, E.C. (1974). Conditional mutator gene in *Escherichia coli*: Isolation, mapping, and effector studies. J. Bacteriol. *117*, 477–487.

Dougan, D.A., Mogk, A., and Bukau, B. (2002). Protein folding and degradation in bacteria: to degrade or not to degrade? That is the question. Cell. Mol. Life Sci. CMLS *59*, 1607–1616.

Duarte, E., Clarke, D., Moya, A., Domingo, E., and Holland, J. (1992). Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc. Natl. Acad. Sci. *89*, 6015–6019.

Eyre-Walker, A., and Keightley, P.D. (2007). The distribution of fitness effects of new mutations. Nat. Rev. Genet. *8*, 610–618.

Fedyunin, I., Lehnhardt, L., Böhmer, N., Kaufmann, P., Zhang, G., and Ignatova, Z. (2012). tRNA concentration fine tunes protein solubility. FEBS Lett. *586*, 3336–3340.

Felsenstein, J. (1974). The evolutionary advantage of recombination. Genetics *78*, 737–756.

Foster, P.L., Lee, H., Popodi, E., Townes, J.P., and Tang, H. (2015). Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. Proc. Natl. Acad. Sci. *112*, E5990–E5999.

Funchain, P., Yeung, A., Stewart, J.L., Lin, R., Slupska, M.M., and Miller, J.H. (2000). The consequences of growth of a mutator strain of *Escherichia coli* as measured by loss of function among multiple gene targets and loss of fitness. Genetics *154*, 959–970.

Gabriel, W., and Bürger, R. (1994). Extinction risk by mutational meltdown: Synergistic effects between population regulation and genetic drift. In Conservation Genetics, D.V. Loeschcke, D.S.K. Jain, and D.J. Tomiuk, eds. (Birkhäuser Basel), pp. 69–84.

Gabriel, W., Lynch, M., and Burger, R. (1993). Muller's ratchet and mutational meltdowns. Evolution. *47,* 1744–1757.

Gustafsson, C., and Persson, B.C. (1998). Identification of the *rrmA* gene encoding the 23S rRNA m1G745 methyltransferase in *Escherichia coli* and characterization of an m1G745-deficient mutant. J. Bacteriol. *180*, 359–365.

Halligan, D.L., and Keightley, P.D. (2009). Spontaneous mutation accumulation studies in evolutionary genetics. Annu. Rev. Ecol. Evol. Syst. *40*, 151–172.

Hershberg, R., and Petrov, D.A. (2010). Evidence that mutation Is universally biased towards AT in bacteria. PLOS Genet *6*, e1001115.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. *44*, D286–D293.

Jaramillo, N., Domingo, E., Muñoz-Egea, M.C., Tabarés, E., and Gadea, I. (2013). Evidence of Muller's ratchet in herpes simplex virus type 1. J. Gen. Virol. *94*, 366–375.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinforma. Oxf. Engl. *28*, 1647–1649.

Keightley, P.D., and Eyre-Walker, A. (2010). What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? Philos. Trans. R. Soc. B Biol. Sci. *365*, 1187–1193.

Kibota, T.T., and Lynch, M. (1996). Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli.* Nature *381*, 694–696.

Kimura, M. (1967). On the evolutionary adjustment of spontaneous mutation rates. Genet. Res. *9*, 23–34.

Kondrashov, F.A., and Kondrashov, A.S. (2010). Measurements of spontaneous rates of mutations in the recent past and the near future. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 1169–1176.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

LeClerc, J.E., Li, B., Payne, W.L., and Cebula, T.A. (1996). High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. Science *274*, 1208–1211.

Lee, H., Popodi, E., Tang, H., and Foster, P.L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc. Natl. Acad. Sci. U. S. A. *109*, E2774-2783.

Lenz, D.H., Mok, K.C., Lilley, B.N., Kulkarni, R.V., Wingreen, N.S., and Bassler, B.L. (2004). The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. Cell *118*, 69–82.

Li, G.-M. (2008). Mechanisms and functions of DNA mismatch repair. Cell Res. *18*, 85–98.

Lind, P.A., and Andersson, D.I. (2008). Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. U. S. A. *105*, 17878–17883.

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res. Notes *5*, 337.

Lindgreen, S., Krogh, A., and Pedersen, J.S. (2014). SNPest: a probabilistic graphical model for estimating genotypes. BMC Res. Notes *7*, 698.

Lynch, M. (2010). Evolution of the mutation rate. Trends Genet. *26*, 345–352.

Lynch, M., and Gabriel, W. (1990). Mutation load and the survival of small populations. Evolution *44*, 1725–1737.

Lynch, M., Conery, J., and Burger, R. (1995). Mutation accumulation and the extinction of small populations. Am. Nat. *146*, 489–518.

Marinus, M.G. (2010). DNA methylation and mutator genes in *Escherichia coli* K-12. Mutat. Res. *705*, 71–76.

Mogk, A., Huber, D., and Bukau, B. (2011). Integrating protein homeostasis strategies in prokaryotes. Cold Spring Harb. Perspect. Biol. *3*, a004366.

Moran, N.A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. U. S. A. *93*, 2873–2878.

Muller, H.J. (1964). The relation of recombination to mutational advance. Mutat. Res. *106*, 2–9.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. *80*, 6278–6281.

Nielsen, R. (2005). Molecular signatures of natural selection. Annu. Rev. Genet. *39*, 197–218.

Rauhut, R., and Klug, G. (1999). mRNA degradation in bacteria. FEMS Microbiol. Rev. *23*, 353–370.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, N. Y.: Cold Spring Harbor Laboratory Pr).

Schoenberg, D.R. (2007). The end defines the means in bacterial mRNA decay. Nat. Chem. Biol. *3*, 535–536.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature *407*, 81–86.

Sittka, A., Pfeiffer, V., Tedin, K., and Vogel, J. (2007). The RNA chaperone Hfq is essential for the virulence of *Salmonella typhimurium*. Mol. Microbiol. *63*, 193–217.

Sledjeski, D.D., Gupta, A., and Gottesman, S. (1996). The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in *Escherichia coli*. EMBO J. *15*, 3993–4000.

Sniegowski, P.D., Gerrish, P.J., and Lenski, R.E. (1997). Evolution of high mutation rates in experimental populations of *E. coli*. Nature *387*, 703–705.

Soll, S.J., Arenas, C.D., and Lehman, N. (2007). Accumulation of deleterious mutations in small abiotic populations of RNA. Genetics *175*, 267–275.

Szamecz, B., Boross, G., Kalapis, D., Kovács, K., Fekete, G., Farkas, Z., Lázár, V., Hrtyan, M., Kemmeren, P., Koerkamp, M.J.A.G., et al. (2014). The genomic landscape of compensatory evolution. PLOS Biol *12*, e1001935.

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P.H., and Godelle, B. (1997). Role of mutator alleles in adaptive evolution. Nature *387*, 700–702.

Tamas, I., Klasson, L., Canbäck, B., Näslund, A.K., Eriksson, A.-S., Wernegreen, J.J., Sandström, J.P., Moran, N.A., and Andersson, S.G.E. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. Science *296*, 2376–2379.

Tamas, I., Wernegreen, J.J., Nystedt, B., Kauppinen, S.N., Darby, A.C., Gomez-Valero, L., Lundin, D., Poole, A.M., and Andersson, S.G.E. (2008). Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. Proc. Natl. Acad. Sci. *105*, 14934–14939.

Tenaillon, O., Barrick, J.E., Ribeck, N., Deatherage, D.E., Blanchard, J.L., Dasgupta, A., Wu, G.C., Wielgoss, S., Cruveiller, S., Médigue, C., et al. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. Nature *536*, 165–170.

Travisano, M., Mongold, J.A., Bennett, A.F., and Lenski, R.E. (1995). Experimental tests of the roles of adaptation, chance, and history in evolution. Science *267*, 87–90.

Wagner, E.G., and Simons, R.W. (1994). Antisense RNA control in bacteria, phages, and plasmids. Annu. Rev. Microbiol. *48*, 713–742.

Wagner, L.A., Weiss, R.B., Driscoll, R., Dunn, D.S., and Gesteland, R.F. (1990). Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli.* Nucleic Acids Res. *18*, 3529–3535.

Wernegreen, J.J., Kauppinen, S.N., and Degnan, P.H. (2010). Slip into something more functional: selection maintains ancient frameshifts in homopolymeric sequences. Mol. Biol. Evol. *27*, 833–839.

Wheeler, N.E., Barquist, L., Kingsley, R.A., and Gardner, P.P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. Bioinforma. Oxf. Engl. *32,* 3566-3574.

Wielgoss, S., Barrick, J.E., Tenaillon, O., Wiser, M.J., Dittmar, W.J., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C., Lenski, R.E., and Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc. Natl. Acad. Sci. *110*, 222–227.

Wohlgemuth, S.E., Gorochowski, T.E., and Roubos, J.A. (2013). Translational sensitivity of the *Escherichia coli* genome to fluctuating tRNA availability. Nucleic Acids Res. gkt602.

Yamazaki, Y., Niki, H., and Kato, J. (2008). Profiling of *Escherichia coli* chromosome database. Methods Mol. Biol. Clifton NJ *416*, 385–389.

Yates, C.M., and Sternberg, M.J.E. (2013). The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein–protein interactions. J. Mol. Biol. *425*, 3949–3963.

Zeyl, C., Mizesko, M., and de Visser, J.A. (2001). Mutational meltdown in laboratory yeast populations. Evol. Int. J. Org. Evol. *55*, 909–917.

CHAPTER 3

Emergence of slippage-prone polyA/T tracts and their impact on fitness

Introduction

RNA editing is broadly defined as the post-transcriptional modification of RNA molecules relative to the corresponding encoding DNA template. Covello and Gray (1993) suggested a general model for the evolution of RNA editing, and the model was later extended by Stoltzfus (1999). In this model, the activity of RNA editing is initiated by pre-existing enzymes. Subsequently, mutation occurs at editable nucleotide positions in the genome and the mutation is fixed in the genome by chance. Upon fixation, RNA editing becomes essential for survival and is maintained by selection. This model for the evolution of RNA editing is an example of a more general scheme of constructive neutral evolution (CNE) that accounts for a complex series of steps giving rise to novel structures and mechanisms (Stoltzfus, 1999). However, this neutral model for the evolution of RNA editing has not been experimentally tested.

Slippage-type editing is a stochastic process analogous to RNA editing and, although it is not a post-transcriptional modification, the general outcome is similar to that of RNA editing, where a change is observed at the phenotypic level and not at the genotypic level. In general, slippage-type editing is the process of RNA polymerase slipping on long polyA/T tracts during transcription and stochastically adding or removing A/Us from the nascent mRNA, resulting in a pool of heterogeneous mRNAs that do not reflect the template (Chamberlin and

Berg, 1962; Tamas et al., 2008; Wagner et al., 1990). The resulting pool of mRNAs may vary: some of these mRNAs are frameshifted, some have codon addition or deletions, and some carry the correct sequence. Therefore, the presence of long polyA/T tracts are generally disadvantageous, and studies on free-living bacterial genomes have shown an underrepresentation of long polyA/T tracts in coding sequences (Ackermann and Chao, 2006; Baranov et al., 2005; Orsi et al., 2010; Sharma et al., 2011), suggesting a strong selection against these slippage-prone tracts. Editing-type processes such as slippage-type editing is therefore a low-fidelity, robust, yet gratuitous, process.

E. coli and *Buchnera* are close relatives and it has been shown that both their RNA polymerases are slippage-prone (Tamas et al., 2008; Wagner et al., 1990), and very few long polyA/T tracts have been observed in the genomes of free-living *E. coli* (Baranov et al., 2005; Orsi et al., 2010). These inherently disadvantageous tracts, however, were observed in abundance within the coding regions (5-50%) of *Buchnera* (Tamas et al., 2008), suggesting an exception to this pattern of selection against slippage-prone tracts. *Buchnera* are inherited maternally and have been through millions of genetic bottlenecks rendering the effect of selection against these slightly deleterious polyA/T tracts ineffective (Mira and Moran, 2002). The maternal transmission enables the fixation of slightly deleterious mutations in the genomes, in a process known as Muller's ratchet (Felsenstein, 1974; Muller, 1964).

In Chapter 2 we subjected *E. coli* populations to single-colony passages, mimicking the genetic background of *Buchnera,* and we observed the accumulation of mutations. Using this mutation accumulation (MA) experiment method, we tested the Covello and Gray's model of the evolution of RNA editing and asked whether slippage-type editing could evolve under lab conditions favouring genetic drift. We then introduced the *araC* gene bearing a frameshifted

polyT tract (observed in our MA experiment) into wild-type *E. coli* and assessed the impact of frameshifted *araC* on cell fitness under conditions where arabinose metabolism is the growth rate limiting factor. Subsequently, we performed RT-PCR and sequencing to examine the consequences of slippage-type editing at the RNA level. We then introduced a bacteriophage antitermination N protein that prevents RNA polymerase slippage (Parks et al., 2014) and examined the effect of the absence of slippage on cell fitness. To our knowledge, this is the first experimental demonstration of the evolutionary drivers for the emergence of a complex RNA editing-like process.

Methods

Strains and media

All chemicals were purchased from Sigma-Aldrich Co. unless otherwise specified. Oligonucleotides were synthesised by Integrated DNA Technologies (USA) and Macrogen (Korea). *Escherichia coli* B strain REL606 was obtained from Tim Cooper (University of Houston, Texas) and REL607 was a spontaneous revertant of REL606 generated in the lab. REL606/7 and REL606/7-derived strains were grown at 37°C in Luria Bertani (LB) media, or Davis Minimal (Difco) supplemented with 1.0 % arabinose (DMA10K), or Davis Minimal without any carbon source (DM0). For solid media, bacteriological agar (Oxoid) was added to a final concentration of 1.5% w/v. All the experiments using *E. coli* were conducted in the presence of antibiotics at the following concentrations: streptomycin, 100μg/mL and ampicillin, 100μg/mL (Peptides International).

The *E. coli B* strain REL606 has a mutation in the *araA* gene that renders it unable to utilise the sugar L-arabinose. Strain REL607 is a spontaneous revertant of REL606 containing a single point mutation that restores the ability to metabolise L-arabinose. The genotypes of all the strains used are defined in Table 3.1.

Table 3.1 | The genotypes of the strains used in this study

Strain	Genotype
REL606	F^-, tsx-467(Am), araA230, lon^-, rpsL227(strR), $hsdR^-$, [mal^+](LamS)
REL606Δ*araC*	F^-, tsx-467(Am), araA230, lon^-, rpsL227(strR), $hsdR^-$, [mal^+](LamS), Δ*araC*
REL606::*araC*	F^-, tsx-467(Am), araA230, lon^-, rpsL227(strR), $hsdR^-$, [mal^+](LamS), insertion of frameshifted *araC*
REL607	F^-, tsx-467(Am), lon^-, rpsL227(strR), $hsdR^-$, [mal^+](LamS)
REL607Δ*araC*	F^-, tsx-467(Am), lon^-, rpsL227(strR), $hsdR^-$, [mal^+](LamS), Δ*araC*
REL607::*araC*	F^-, tsx-467(Am), lon^-, rpsL227(strR), $hsdR^-$, [mal^+](LamS), insertion of frameshifted *araC*

Experimental evolution

A total of 10 genetically identical lineages were derived from a single glycerol stock of REL606 that contains a pGEM::*mutD5* plasmid. Single colonies of each individual line were randomly picked, streaked onto fresh Luria Bertani (LB) agar and grown at 37˚C for 24 hours. The pick-streak-incubate process (growth cycle) was repeated for 100 growth cycles, where the cells were maintained on LB agar plates supplemented with streptomycin and ampicillin. Five non-bottlenecked control lines were also prepared in parallel where the cells were washed with 1 mL of 1xPBS and 100 uL of the washed cells were transferred and streaked onto a fresh LB agar plate (see Chapter 2). Approximately 10^7 to 10^8 of washed cells were transferred for the non-bottleneck control lines while the number of cells transferred for the bottlenecked line changed every day as the colonies were randomly selected (approximately 10^3). A glycerol stock of each lineage was prepared every day (bottlenecked) and every ten passages (non-bottleneck), generating a fossil record.

Genome sequencing and mapping

The bottlenecked and non-bottlenecked lineages, at intervals of 10 days, were streaked to single colonies on LB agar. Single colonies were then used to inoculate LB liquid media supplemented with streptomycin and ampicillin. Genomic DNA was isolated using the Wizard Genomic DNA Purification Kit (Promega) and quantified using the Nanodrop 1000 Spectrophotometer and Qubit 2.0 Fluorometer. Sequencing was carried out by Macrogen Korea using the Illumina MiSeq platform with 2x250bp paired-end reads. The resulting sequencing reads were processed with two different pipelines. In the first pipeline, AdapterRemoval (Lindgreen, 2012) was used to remove low-quality reads and adapter sequences. Reads were then mapped to the REL606 genome (NC_012967) using Bowtie2 (Langmead and Salzberg, 2012) (using default parameters, specifying haploid genomes where necessary), and genotyping was carried out using SNPest (Lindgreen et al., 2014). In the second pipeline, the raw reads were processed in Geneious v9.1.3 (Kearse et al., 2012), the adaptors were removed using the BBDuk plug-in (Bushnell 2014). The trimmed reads were then mapped to the REL606 genome using the Bowtie2 plug-in for Geneious (Langmead and Salzberg, 2012), and SNP-calling was performed. The use of two pipelines enables us to get the most in-depth analysis of the genome as the stringency of the SNP calling varies between the software with regards to the read depth.

Scarless allelic replacement

A suicide plasmid-based scarless allele replacement method utilising homologous recombination (Fehér et al., 2008) was used to replace the *araC* gene of wild-type REL606 and REL607, which bears an in-frame poly 7T tract, with the *araC* bearing a frameshifted poly 8T tract. Briefly, an 1.1 kb-long targeting DNA fragment carrying the desired point mutation in the middle of *araC* was amplified using Phusion® High-Fidelity DNA

polymerase with primers flanking the point mutation (approximately 500 bp upstream and downstream of the mutation). These primers were designed to consist of restriction sites at the 5' end enabling efficient cloning (Table 3.2, sequence is indicated in blue). The PCR products were purified (Promega), restricted with SacI (Fermentas) and XmaI (NEB) and cloned into the temperature-sensitive suicide plasmid, pST76-A. The plasmid construct, pST76-A::*araC* was used to transform the *E. coli* DH5α strain. The pST76-A::*araC* plasmid was then purified utilising the PureLink® Quick Plasmid Miniprep Kit (Invitrogen) and transformed into *E. coli* REL606. The resulting transformants were grown at a temperature non-permissible for plasmid replication (42°C), resulting in cells with the plasmid integrated into the chromosome via a single crossover between the mutant allele and the corresponding chromosomal region. The cells that screened positive for the integrated plasmid were then transformed with the pSTKST helper plasmid which carries the gene that encodes I-SceI. Recombination was induced with chlortetracycline-hydrochloride (CTC) resulting in the expression of I-SceI meganuclease enzyme, which cleaves the chromosome at the 18-bp recognition site in the integrated suicide plasmid. RecA-mediated intramolecular recombination repairs the chromosomal gap via the broken ends carrying short homologous regions resulting in either the original wild-type sequence (*araC* 7T) or the replacement gene (*araC* 8T). The sequence flanking the gene of interest was sequenced bi-directionally to confirm knock-ins.

The same scarless allelic replacement method was used to knock-out *araC* from REL606 and REL607 (Figure 3.1). However, instead of directly cloning the gene carrying the point mutation into pST76-A, the regions flanking *araC* (~500 bp upstream and downstream) were amplified (using primers d and br, and e and bf) and subsequently, the PCR products were combined and amplified using primers a and c, which consist of restriction sites for efficient

cloning. The amplified and digested PCR product that spans the junction of *araC* gene was subsequently cloned into the suicide pST76-A plasmid for plasmid integration followed by recombination with the helper, pSTKST, as described above. The primers utilised for *araC* knock-in and knock-out are listed in Table 3.2.

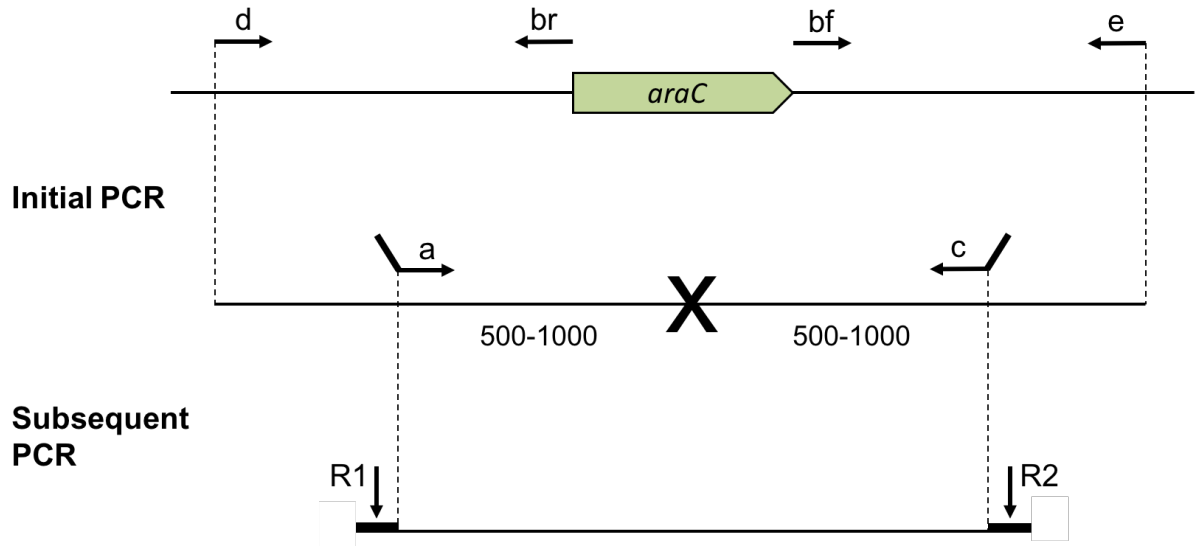

Figure 3.1 | The scheme for generating *araC* knock-out. The numbers indicate the suggested length of the nucleotide sequences and the arrows and lower case alphabet indicate the primers. Primers br and bf span the knockout junction and are reverse complements of each other at the 5′ end. Image adapted from Fehér et al., 2008.

Table 3.2 | Primers used in the scarless allelic replacement

Primer	Sequence 5′-3′
araC_KI_a	GATCGAGCTCGTTATTCGGGGCATCACAAAATTGC
araC_KI_c	GATCCCCGGGCGGAAAAGATGTGACAGACGCGACG
araC_KI_d	CAATGTAGTCACGCGGATGATGACG
araC_KI_e	CACCACCGCCATCAATGAATACACG
araC_KO_a	GATCGAGCTCAAACGGGTAATCTCTTCCGCTTC
araC_KO_c	GATCCCCGGGACTTAAACGATCGGTTGCTTTACC
araC_KO_d	GAATAAATACCGCCAATATAGC
araC_KO_e	GCAAAATATCGATATACACCGGC
araC_KO_br	CAATTGTCTGATTCGTTACCAAACTTTTCATACTCCCGCCATTC
araC_KO_bf	GAATGGCGGGAGTATGAAAAGTTTGGTAACGAATCAGACAATTG
pST76-A t1	CGGAAGGATCTGAGGTTCTTATGGC
pST76-A t2	CGAATTGTCGACAAGCTTGATCTGGC

The abbreviations KI and KO represent knock-in and knock-out respectively. The KI primers were used for frameshifted *araC* knock-in while the KO primers were used for *araC* knock-out. The blue sequences represent restriction enzyme cleavage sites.

Selection for arabinose revertants

The *E. coli* B strain REL606 is an Ara$^-$ strain with a mutation in the *araA* gene that prevents the utilisation of the sugar L-arabinose. Strain REL607, however, is a spontaneous revertant of REL606 containing a single point mutation in the *araA* gene (A275G) that restores the ability to metabolise L-arabinose. This marker is selectively neutral under conditions where arabinose is not the sole carbon source (Maddamsetti et al., 2015). Instead of performing a scarless replacement of *araC* in REL607 we reverted our REL606::*araC,* an Ara$^-$ strain, to REL607::*araC,* an Ara$^+$ strain. REL606::*araC* was revived and grown in 5 mL of Davis Minimal supplemented with 0.1% glucose (DM1000) and streptomycin at 37°C overnight, reaching an approximate density of 5 x 10^9 cell/mL. Approximately 500 cells were inoculated

into 10 mL of DM1000 and the cultures were incubated at 37°C overnight. The culture was spun down and plated on Minimal Arabinose (MA) solid plate where arabinose is the sole carbon source. The plate was incubated at 37°C for 48 hours and single colonies were picked and streaked onto new MA plates to avoid false positives. Subsequently, the colonies were screened for Ara^+ mutants via PCR-RFLP.

PCR-RFLP assay (Ara^+ mutation verification)

PCR using KAPA2G Robust HotStart ReadyMix PCR Kit (KAPA Biosystems) with REL256 and REL257 primer pair was performed on the whole REL607::*araC* cells from the above MA plate, producing a 495-bp product. A proportion of the PCR products was then subjected to restriction digest with HaeIII. An Ara^- strain would result in 3 bands (226, 197 and 72-bp) while an Ara^+ strain would result in 4 bands (207, 197, 72 and 19-bp). The undigested PCR products amplified with REL256 and REL257 which screened positive as REL607 (Ara^+ strain) were purified and sequenced bi-directionally (primer sequences were obtained from http://barricklab.org/twiki/bin/view/Lab/WebHome).

Primers used to amplify *araA* were:

REL256:

5′ – CCGATACGCTCATGGGCTTGTTTA – 3′

REL257:

5′ – CTGCCCAGGCCGTTGCGACTCTAT – 3′

Total RNA extraction and RT-PCR

The REL606::*araC* and REL607::*araC* strains were grown in LB supplemented with streptomycin at 37°C. The bacterial cultures were diluted 1:100 in 10 mL of fresh LB supplemented with streptomycin. The cultures were grown for approximately 3 hours and total RNA was isolated from the mid-log phase cultures using a hot phenol method (Schmitt et al., 1990). Purified total RNA was diluted to 300 ng/µL and treated with TURBO DNase I (Ambion) following the manufacturers' guidelines. The DNase I-treated total RNA was amplified using Phusion® High-fidelity DNA Polymerase (Thermo Fisher) with araC_KI_F and araC_KI_R primers to check for genomic DNA contamination. The DNA-free total RNA was then subjected to RT-PCR using the SuperScript® III One-Step RT-PCR System with Platinum® Taq DNA Polymerase kit (Invitrogen) with gene specific primers: araC_KI_F and araC_KI_R following the manufacturers' guidelines.

Primers used to detect *araC* were:

araC_KI_F:

5′ – GATGCAATATGGACAATTGGTTTC – 3′

araC_KI_R:

5′ – ATGACGACCGTAGTGATGAATCTC – 3′

The REL607::*araC* strain was grown in LB supplemented with streptomycin at 37°C. The bacterial culture was diluted 1:100 in 10 mL of fresh Davis minimal supplemented with 1% arabinose (DMA10K) and streptomycin. The culture was grown for approximately 6 hours and total RNA was isolated, DNase treated and checked for genomic DNA contamination as previously described. The DNA-free total RNA was then subjected to first-strand synthesis using SuperScript II (Invitrogen) with gene specific primers: araC_KI_F and araC_KI_R

following the manufacturers' guidelines. The first-strand cDNA was then amplified using Phusion® High-fidelity DNA Polymerase (Thermo Fisher) with araC_KI_F and araC_KI_R. The amplified PCR products were then A-tailed with Taq Polymerase (Bioline), cloned into pGEM-T easy (Promega) and sequenced bi-directionally (Macrogen). The total RNA was also checked for mRNA viability prior to first-strand synthesis with *gstA* specific primers using SuperScript® One-Step RT-PCR System with Platinum® *Taq* DNA Polymerase kit (Invitrogen).

Primers used to detect *gstA* were:

gstA_fwd:

5′ – CTTTGCCGTTAACCCTAAGGG – 3′

gstA_rev:

5′ – GCTGCAATGTGCTCTAACCC – 3′

AraC structural modelling

The mutated *araC* sequence was used to create a homology model with SWISS-MODEL (Biasini et al., 2014). The *E. coli* AraC regulatory protein structure complexed with L-arabinose, 2ARC (Soisson et al., 1997) was used as the template for the modelling our frameshifted AraC. Protein structures were visualized using PyMOL v1.3r1 (Schrödinger, 2015).

Optical density and growth rate measurement

The REL606 and REL607 strains were streaked to single colonies and these single colonies were grown in LB supplemented with streptomycin at 37°C for 2 days. The bacterial cultures were diluted 1:100 in 0.1 L of either fresh; 1) LB supplemented with streptomycin or; 2) Davis Minimal supplemented with 1% arabinose (DMA10K) and streptomycin in a 24 well cell culture plate. The optical density (OD595) was measured with an FLUOstar Omega Microplate Reader (BMG Labtech) at 37°C (with shaking at 200 rpm) for 16-24 hours. The OD595 was measured every 6 min, with the plate shaking between readings. All experiments were performed with a minimum of 3 biological replicates, along with 3 technical replicates. The growth rates for each line (averaged across replicate cultures) were determined as the minimum doubling time taken over a 30-minute interval.

Introduction of N protein

A plasmid bearing the N protein (pNAS150), kindly donated by Donald Court (NIH), was used to transform the REL607 and REL607::*araC* strains using calcium chloride and heat shock (Sambrook et al., 1989), generating REL607 + pNAS150 and REL607::*araC* + pNAS150 strains. The growth of rates REL607 + pNAS150 and REL607::*araC* + pNAS150 were measured in Davis Minimal supplemented with 1% arabinose (DMA10K) using a plate reader as described in the method above (Optical density and growth rate measurement).

Results

Emergence of frameshifted polyA/T under bottleneck conditions

To test whether genetic drift plays a role in driving the emergence of transcriptional slippage, we subjected 10 individual *E. coli* lineages to serial single-colony passages and 5 lineages to non-bottlenecked passages. The genomes of all the lineages were sequenced after 50 and 100 passages. A total of 22 frameshifted polyA/T tracts (n≥9) emerged in the coding region bottlenecked lineages and 1 frameshifted polyA/T tracts (n≥9) in the non-bottlenecked lineages after 100 serial passages (Figure 3.2 and Table 3.3).

The genes bearing these frameshifted polyA/T tracts (Table 3.3) were considered as non-essential according to the Profiling of *E. coli* Chromosome (PEC) database (Yamazaki et al., 2008). According to PEC, essential genes are those genes that are essential for cell growth while non-essential genes are genes that are dispensable for cell growth under the conditions of the experiment. The mutations that generated frameshifted polyA/T tracts were mostly enriched in genes that code for hypothetical proteins. However, some of the genes that contain these polyA/T tracts were documented to play a role in information storage, cellular processes, and metabolism (Supplementary, Figure 3.1). It is important to note that we observed the emergence of a frameshifted poly 8T tract within the *araC* gene.

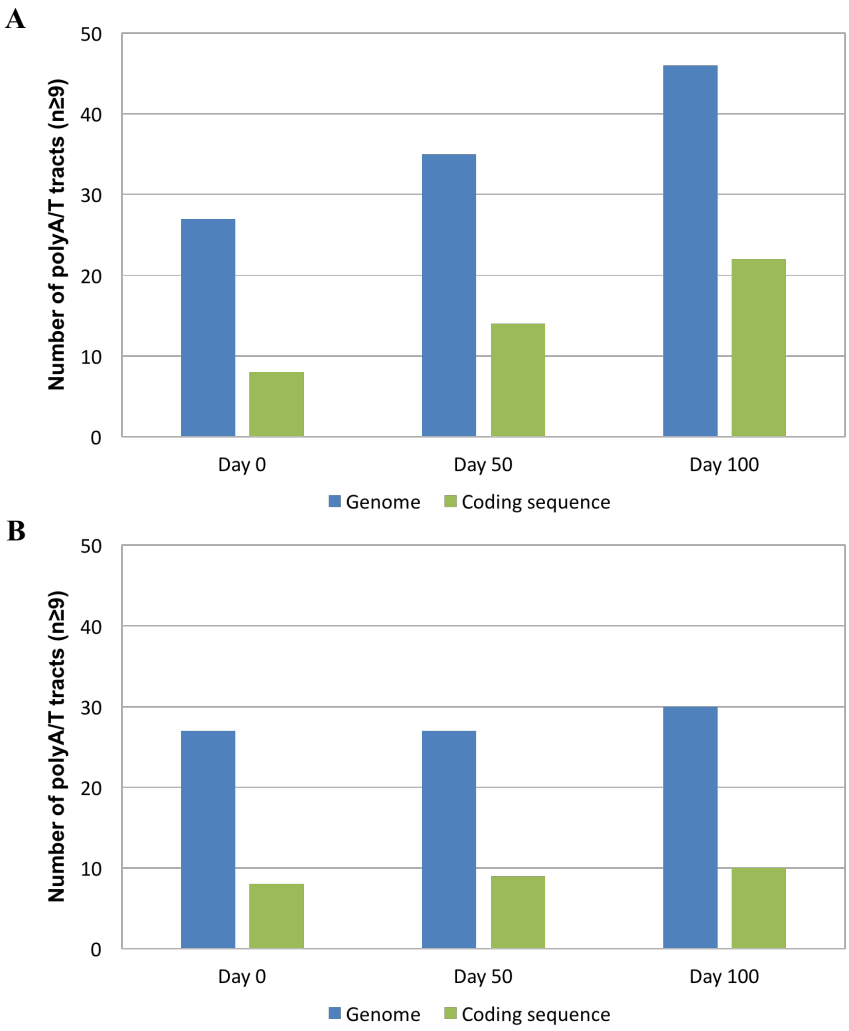

Figure 3.2 | Higher numbers of polyA/T tracts (n≥9) were observed in the bottlenecked lineages compared to the non-bottlenecked control lineages. Blue bars represent the number of polyA/T tracts in the genome and the green bars represent the number of frameshifted polyA/T tracts in the coding sequences. **A)** The number of polyA/T tracts observed in the bottlenecked lineages. **B)** The number of frameshifted polyA/T tracts observed in the control lineages.

Table 3.3 | Genes bearing frameshifted polyA/T tract (n≥9)

Locus tag	Gene	DNA change	Protein change	Lineage with mutation	Gene function
ECB_00377		(A)8 to (A)9	Frameshift	BN100.1	Putative acetyltransferase
ECB_00400	*mdlA*	(A)8 to (A)9	Frameshift	BN100.10	Putative multidrug ABC transporter ATPase (Allikmets et al., 1993)
ECB_00535	*cusB*	(A)7 to (A)9	Frameshift	BN100.1	Protein transporter activity (Franke et al., 2003)
ECB_00652	*kdpD*	(T)8 to (T)9	Frameshift	BN100.1	Two-component sensor activity (Walderhaug et al., 1992)
ECB_00775	*ybiO*	(T)8 to (T)9	Frameshift	BN100.1	Plasma membrane (Edwards et al., 2012)
ECB_00831		(A)8 to (A)9	Frameshift	BN100.1 and BN100.9	Retron DNA polymerase-like protein
ECB_00852	*ybjL*	(A)8 to (A)9	Frameshift	BN100.6	Predicted transporter
ECB_01154	*ycgL*	(A)8 to (A)9	Frameshift	BN100.6	Unknown function
ECB_01341		(A)8 to (A)9	Frameshift	BN100.6	Hypothetical protein
ECB_01396	*ydcR*	(A)8 to (A)9	Frameshift	BN100.1 and BN100.8	Uncharacterized HTH-type transcriptional regulator and predicted amino transferase
ECB_01528		(T)8 to (T)9	Frameshift	BN100.1	Antitermination protein Q-like protein
ECB_01632	*cfa*	(T)9 to (T)10	Frameshift	BN100.1	Unsaturated-phospholipid methyltransferase (Taylor et al., 1981)
ECB_01747	*yeaA*	(T)8 to (T)9	Frameshift	BN100.1	Peptide-methionine-(S)-S-oxide reductase activity (Ezraty et al., 2005)
ECB_02357	*ypfG*	(A)8 to (A)9	Frameshift	BN100.1	Uncharacterized protein
ECB_02500	*ypjD*	(A)7 to (A)9	Frameshift	BN100.1	Predicted inner membrane protein
ECB_02652		(T)8 to (T)9	Frameshift	BN100.1	Hypothetical protein
ECB_02798		(A)9 to (A)10	Frameshift	BN100.6	Hypothetical protein
ECB_02821		(A)9 to (A)10	Frameshift	C100.1	Hypothetical protein
ECB_03398	*yhjY*	(T)8 to (T)9	Frameshift	BN100.1	Putative outer membrane protein
ECB_03973	*phnG*	(A)7 to (A)9	Frameshift	BN100.1	Phosphonate transport (Metcalf and Wanner, 1991)
ECB_04004	*cadC*	(A)8 to (A)9	Frameshift	BN100.8	Transcriptional activator (Watson et al., 1992)
ECB_04141	*yjhB*	(T)8 to (T)9	Frameshift	BN100.1	Putative transporter

BN represents the bottlenecked lineages while C represents the non-bottlenecked (control) lineages. The numbers following BN and C denote the number of passages followed by the lineage number respectively.

Scarless allelic replacement produced araC knock-outs and knock-ins

To assess the impact of frameshifted *araC* (observed in our evolution experiment) on cell fitness, *araC* was knocked-out (generating REL606Δ*araC* and REL607Δ*araC*) and the frameshifted *araC* was knocked in separately (generating REL606::*araC* and REL607::*araC*) using the scarless genome engineering protocol (see methods). REL606Δ*araC* and REL607Δ*araC* (Figure 3.3), and REL606::*araC* and REL607::*araC* (Figure 3.4) strains were obtained and confirmed by bi-directional sequencing.

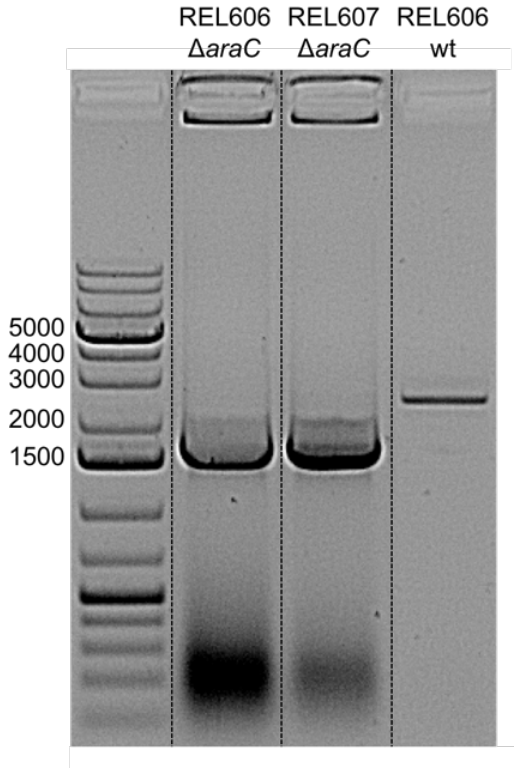

Figure 3.3 | The *araC* gene was successfully knocked-out from wild-type REL606 and REL607. The colonies were screened with primers flanking the *araC* gene prior to confirmation of knock-out by sequencing. Lane 1, GeneRuler 1kb Plus (Thermo Scientific) with size standards indicated. Lane 2, PCR product for REL606Δ*araC* (expected size, 1619 bp), lane 3, PCR product for REL607Δ*araC* (expected size, 1619 bp), and lane 4, PCR product for wild-type REL606 bearing the *araC* gene (expected size, 2498 bp).

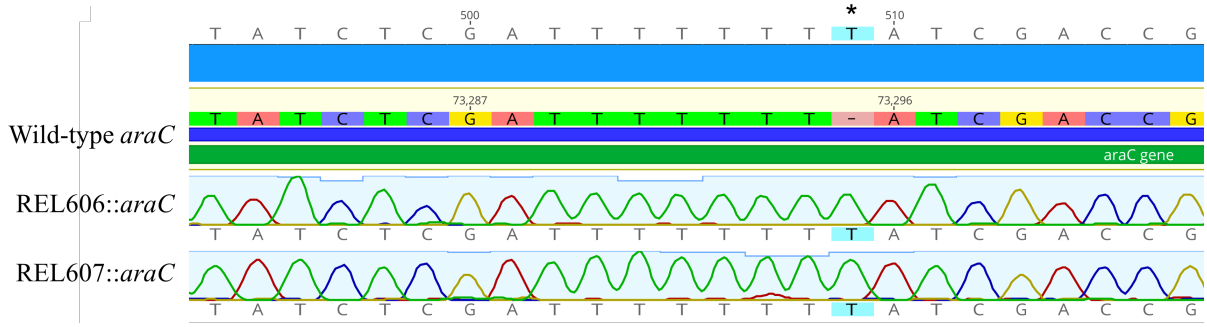

Figure 3.4 | Sequencing of the *araC* gene revealed a successful knock in of the frameshifted *araC* gene. A single base insertion of a thymine residue (indicated by an asterisk) was observed in REL606::*araC* and REL607::*araC* but not in wild-type REL606.

A frameshift mutation in araC produces truncated AraC protein

To assess the impact of the observed frameshift mutation on AraC structure, we created a homology model of AraC using the structure for *E. coli* AraC complexed with L-arabinose, 2ARC (Soisson et al., 1997). We observed a truncated AraC structure that could not form a homodimer (Figure 3.5).

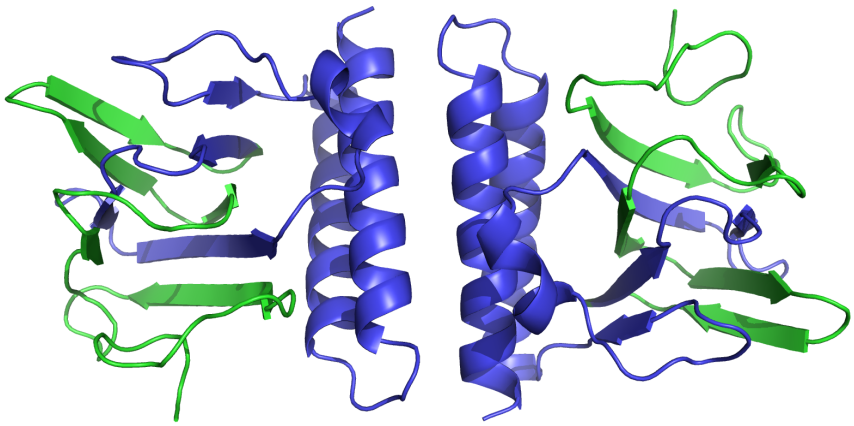

Figure 3.5 | The crystal structure of *E. coli* regulatory homodimer AraC protein (2ARC). The green loop and beta sheets represent the region of AraC with the observed frameshift mutation that was mapped onto 2ARC forming truncated AraC monomers. The truncated AraC structure does not consist of the alpha helices (blue regions) required for dimerisation of the monomers.

RT-PCR confirmed the production of araC transcripts

To assess whether the knocked in *araC* gene is producing RNA transcripts, RT-PCR was conducted using primers internal to the *araC* mRNA. Transcripts resulting from the introduction of *araC* were detectable in both REL606::*araC* and REL607::*araC* reaffirming successful knock-in of the *araC* gene (Figure 3.6).

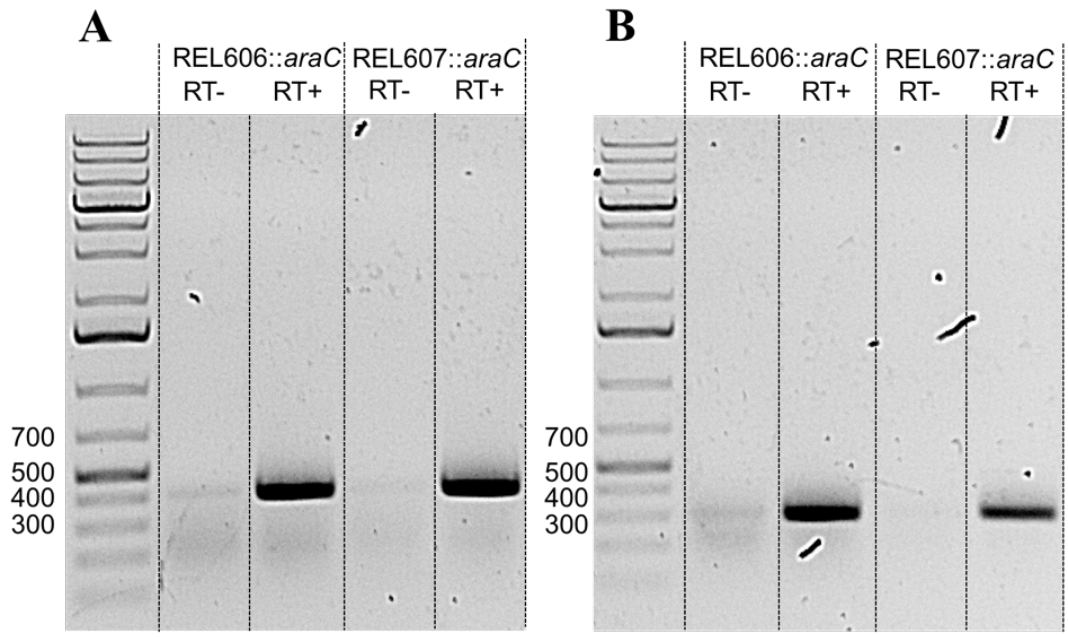

Figure 3.6 | RT-PCR confirms the expression of *araC* transcripts in both REL606::*araC* and REL607::*araC*. Total RNA was extracted from mid-log phase cultures of REL606::*araC* and REL607::*araC*, and were used as a template for RT-PCR. RT- lanes were processed without a reverse transcriptase reaction in order to detect the presence of genomic DNA. RT+ lanes were processed as per the manufacturer's guidelines that included a reverse transcription step. **A)** RT-PCR with primers corresponding to REL606 *gstA* gene (expected size 411 bp). Lane 1, GeneRuler 1 kb Plus (Thermo Scientific) with size standards indicated; lanes 2-5, RT-PCR products for REL606::*araC* and REL607::*araC* with RT- controls indicated. **B)** RT-PCR with primers corresponding to *araC* (expected product size 304 bp). Lane 1, GeneRuler 1 kb Plus (Thermo Scientific) with size standards indicated; lanes 2-5, RT-PCR products for REL606::*araC* and REL607::*araC* with RT- controls indicated.

REL606 and REL607 araC knockouts do not express araC

Although we have shown successful knock-out of the *araC* genes from REL606 and REL607 by sequencing, to ensure that the *araC* gene did not recombine into other loci within the chromosome, we checked for the expression of *araC* by RT-PCR. Primers internal to *araC* and *gstA* mRNA were used. Transcripts resulting from the housekeeping gene *gstA* gene were detectable but not the *araC* gene in both REL606Δ*araC* and REL607Δ*araC*, confirming successful knock-out of the *araC* gene from the REL606 and REL607 chromosomes (Figure 3.7).

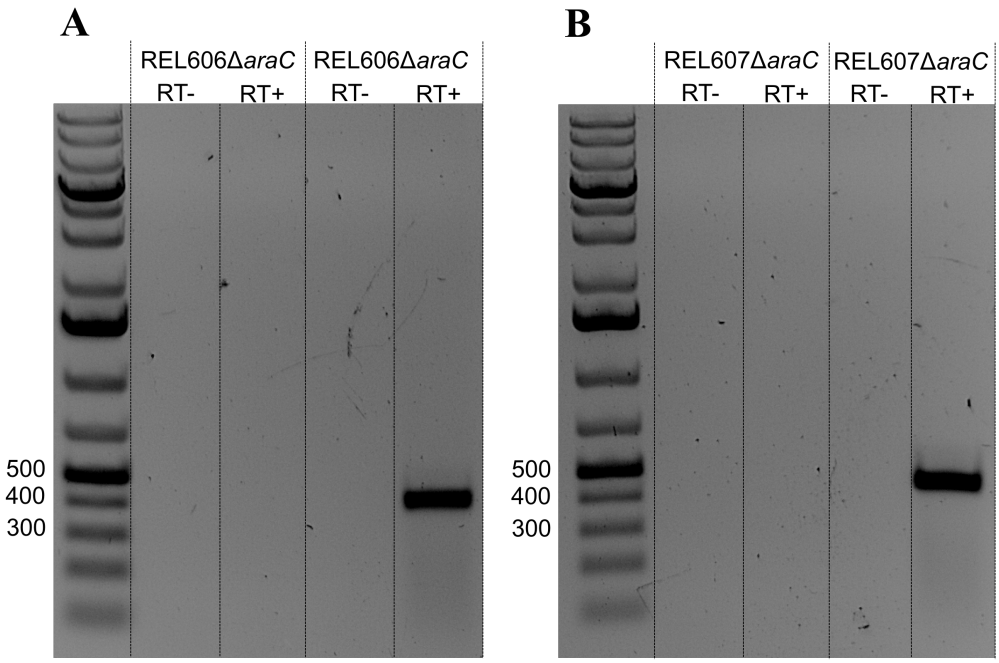

Figure 3.7 | RT-PCR confirms the absence of *araC* transcript expression in both REL606Δ*araC* and REL607Δ*araC*. Total RNA was extracted from mid-log phase cultures of REL606Δ*araC* and REL607Δ*araC* and was used as a template for RT-PCR. RT- lanes were processed without a reverse transcriptase reaction in order to detect the presence of genomic DNA. RT+ lanes were processed as per the manufacturer's guidelines, that included a reverse transcription step. **A)** RT-PCR with primers corresponding to REL606 *araC* and *gstA* gene (expected size 304 bp and 411 bp respectively). Lane 1, GeneRuler 1 kb Plus

(Thermo Scientific) with size standards indicated; lanes 2-3, RT-PCR products for REL606Δ*araC* amplified with *araC* primers and lanes 4-5, RT-PCR products for REL606Δ*araC* amplified with *gstA* primers with RT- controls indicated. **B)** RT-PCR with primers corresponding to REL606 *araC* and *gstA* genes (expected size 304 bp and 411 bp respectively). Lane 1, GeneRuler 1 kb Plus (Thermo Scientific) with size standards indicated; lanes 2-3, RT-PCR products for REL607Δ*araC* amplified with *araC* primers and lanes 4-5, RT-PCR products for REL607Δ*araC* amplified with *gstA* primers with RT- controls indicated.

AraC is not essential in rich media

We observed an addition of a thymine nucleotide to the *araC* gene (change from in-frame poly 7T to a frameshifted poly 8T) among 4,000 other SNPs in one of our bottlenecked lines (BN100.1) during our experimental evolution (discussed in Chapter 2 and above). To examine the impact of the frameshifted *araC* on cell fitness, we knocked out *araC* and knocked in the frameshifted *araC* gene in REL606 and REL607. Subsequently, we measured the growth rates in LB media to test the effect of *araC* on fitness (Figure 3.8). The calculated minimum doubling times of REL606Δ*araC* and REL606::*araC* were not significantly different at 25.0±1.2 min and 26.9±1.0 min, respectively compared to wild-type REL606 doubling time of 26.5±0.7 min (*P*-value = 0.196, Kruskal-Wallis, $\bar{X}$±SE, n=6). We also measured the growth rates of the REL607-derived strains and there was no significant difference between the doubling times of REL607Δ*araC* and REL607::*araC* of 24.9±0.7 min and 26.0±0.2 min, respectively compared to the wild-type REL607 doubling time of 25.5±0.7 min (*P*-value = 0.459, Kruskal-Wallis, $\bar{X}$±SE, n=6). Additionally, there were no significant differences between all the REL606 and REL607 strains tested (*P*-value = 0.576, Kruskal-Wallis, $\bar{X}$±SE, n=6).

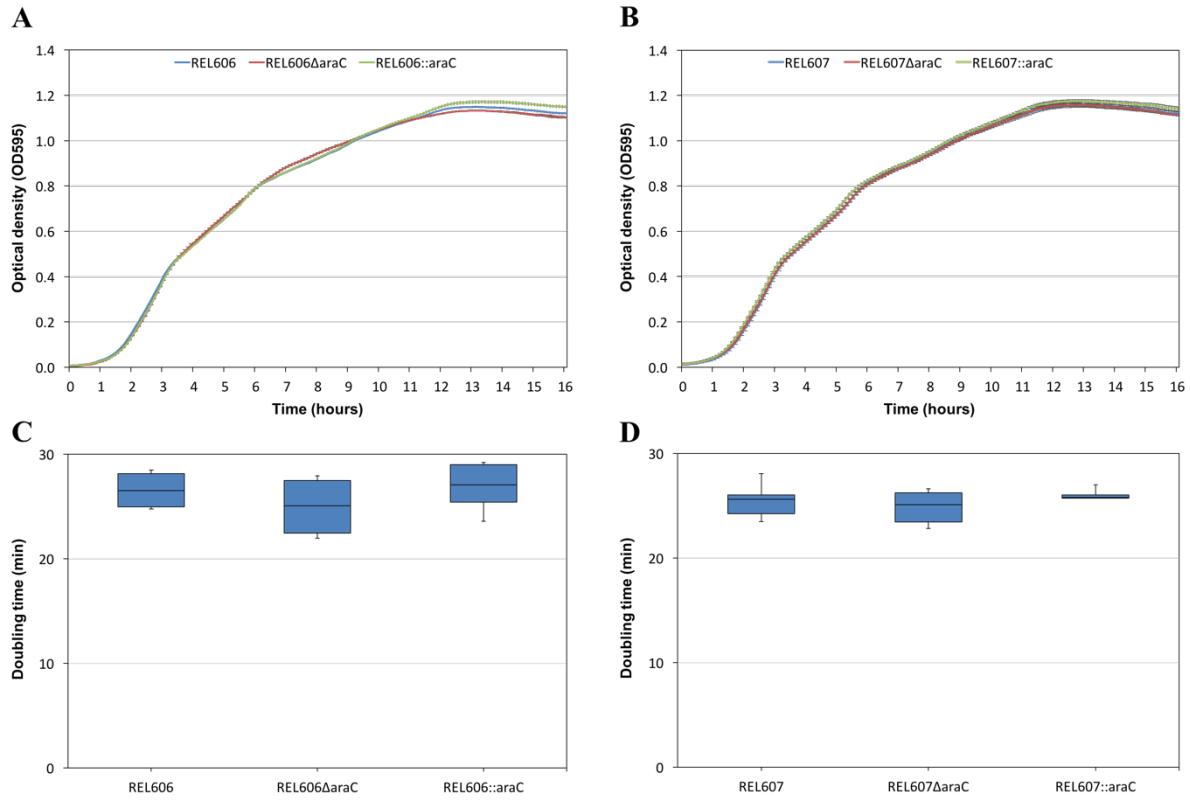

Figure 3.8 | There was no significant difference in growth rates between *E. coli* REL606 and REL607 and their derived strains in rich LB media. Three REL606/7 (blue), REL606/7Δ*araC* (red), and REL606/7::*araC* (green) clones were grown in LB media at 37°C with shaking, and the optical density (OD595) was collected every 6 minutes for 16 hours. The OD595 points are shown as $\bar{X}$±SD, n=6. **A)** The OD595 of REL606 and its derived strains over a 16-hour period. **B)** The OD595 of REL607 and its derived strains over a period of 16 hours. **C)** The doubling times of REL606, REL606Δ*araC*, and REL606::*araC* were 26.5±0.7 min, 25.0±1.1 min, and 26.9±1.0 min respectively. **D)** The doubling times of REL607, REL607Δ*araC*, and REL607::*araC* were 25.5±0.7 min, 24.9±0.7 min, and 26.0±0.2 min respectively. Doubling times are shown as $\bar{X}$±SE, n=6. There was no observable difference in the growth rates among any of the strains tested (*P*-value = 0.576, Kruskal-Wallis, $\bar{X}$±SE, n=6).

Cell fitness decreased upon the introduction of frameshifted araC in minimal media supplemented with arabinose

Under conditions of rich LB media containing glucose, we did not see any significant difference in doubling time between wild-type REL606 and REL607 compared to their derived strains (Figure 3.8). To assess the impact of *araC* (*araC* knock-out and frameshifted *araC* knock-ins) on cell fitness, we measured the growth rates in Davis minimal supplemented with 1% arabinose as the sole carbon source (DMA10K). Three colonies of REL606/7 and REL606/7Δ*araC* were grown to saturation in LB media before diluting 1:100 in fresh DMA10K to monitor growth. The cells were not washed of LB prior to the dilution, but the amount of LB was only sufficient to maintain the growth of a small number of cells, as observed when the cells were grown in just Davis minimal media without any form of sugar (discussed below).

No growth was observed in minimal media supplemented with arabinose only for all the REL606 strains and REL607Δ*araC* (Figure 3.9). Growth was observed for REL607::*araC* and REL607 where a decreased in growth rate was observed for REL607::*araC* compared to wild-type REL607. The calculated doubling time for the REL607::*araC* isolate increased to 90.3±1.0 min from 61.0±0.4 min calculated for the wild-type REL607 (*P*-value = 7.40E-07, Wilcoxon rank sum test, $\bar{X}$±SE, n=12). In addition, a longer lag phase was also observed in REL607::*araC* compared to REL607.

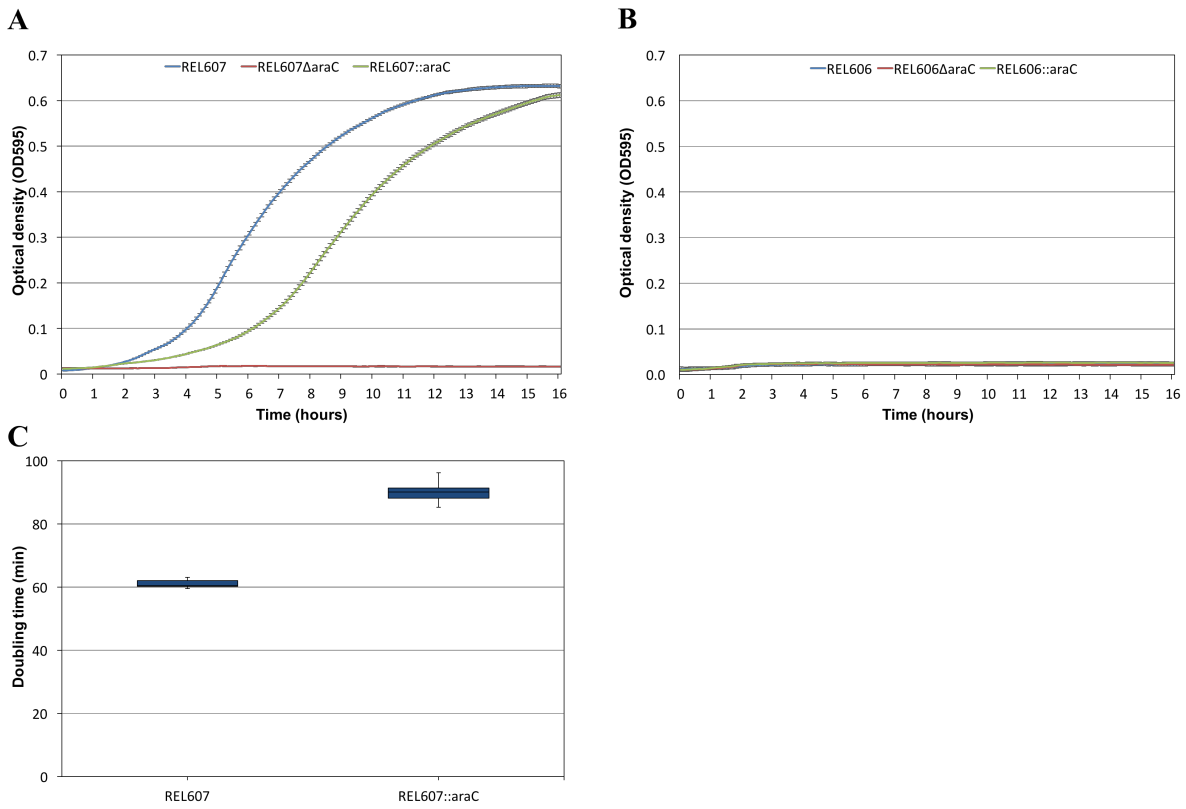

Figure 3.9 | *E. coli* REL606, REL606-derived, and REL607Δ*araC* do not grow in minimal media where arabinose is the sole carbon source and the introduction of frameshifted *araC* into REL607 decreases the growth rate. Three REL606/7 (blue), REL606/7Δ*araC* (red), and REL606/7::*araC* (green) clones were grown in Davis minimal supplemented with 1% arabinose (DMA10K) at 37°C with shaking and the optical density (OD595) was collected every 6 minutes for 16 hours. The OD595 points are shown as $\bar{X}$±SD, n=12. **A)** The OD595 of REL607 and its derived strains were monitored for 16 hours. **B)** The OD595of REL606 and its derived strains were monitored for 16 hours. **C)** The doubling time of REL607, and REL607::*araC* were 61±0.3 min and 90±1.0 min respectively. Doubling times are shown as $\bar{X}$±SE, n=12. The calculated doubling time increased with the introduction of frameshifted *araC* in DMA10K (*P*-value = 7.40E-07, Wilcoxon rank sum test, $\bar{X}$±SE, n=12).

To test whether the LB from the overnight cultures affects growth in arabinose only media, we grew the cells in LB overnight, diluted the culture 1:100 into Davis minimal with no carbon source (DM0) and measured the growth rates. A slight increase in optical density from 0.015 to 0.020 was observed for both REL606 and REL607 but the optical density plateaued after 3 hours (Figure 3.10).

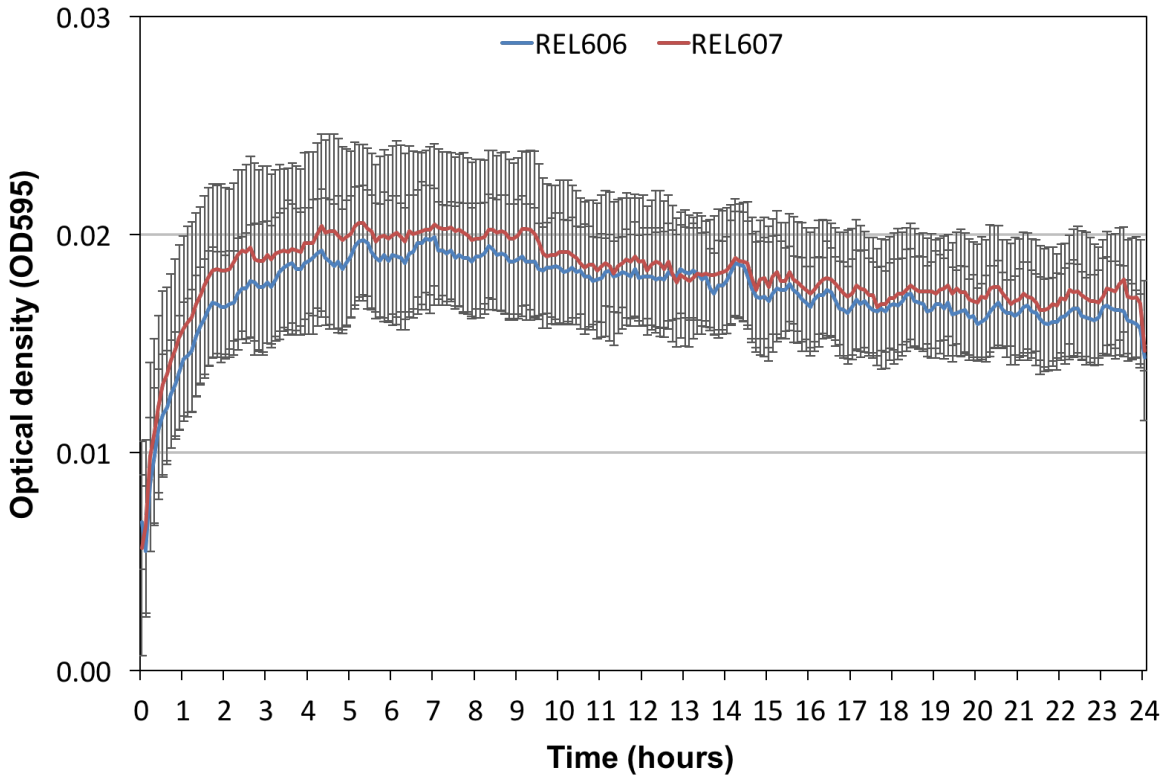

Figure 3.10 | Cells do not reach saturation in minimal media (DM0) in the absence of a carbon source. The overnight cultures grown in rich media were diluted 1:100 in DM0 and the growth rate was monitored every 6 minutes for 24 hours. A rise in OD595 was observed between 0-3 hours and the readings plateaued beyond the 3-hour mark. The OD595 points are shown as $\bar{X}$±SD, n=3.

A heterogeneous population of mRNA was observed

To assess the impact of the introduction of a frameshifted poly 8T tract into the *araC* gene, we grew the cells in Davis minimal supplemented with 1% arabinose (DMA10K) as the sole carbon source. Subsequently, total RNA was extracted at mid-log phase, reverse transcribed, PCR amplified, cloned and 100 clones were sequenced. Our results show that poly 8T tracts are prone to RNA polymerase slippage and heterogeneity in the length of the polyT tracts was observed, where a proportion of transcripts were in the correct frame (Figure 3.11). The percentage of RNA polymerase slippage on the poly 8T tract is approximately 4.8%.

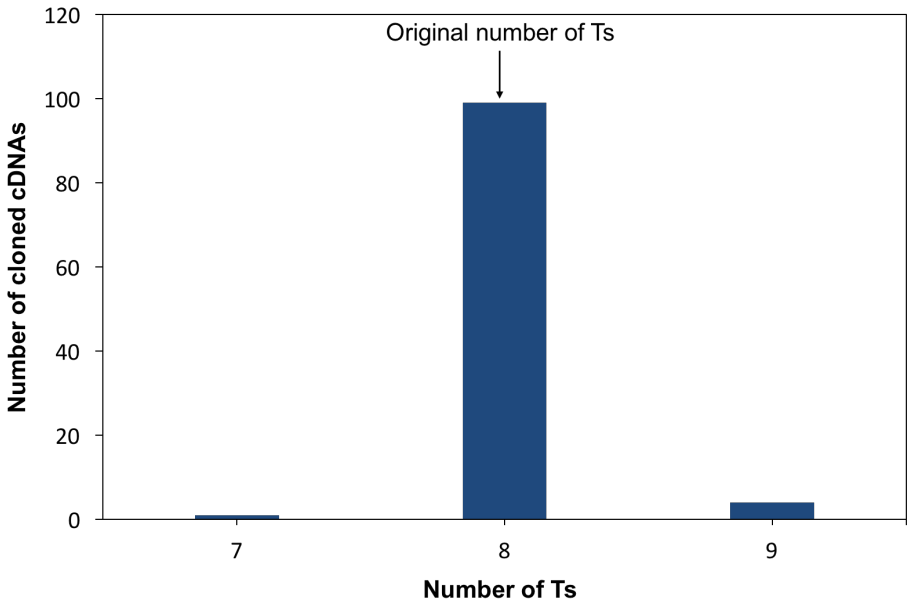

Figure 3.11 | Slippage resulted in transcript with varying lengths. Although, low levels of slippage were observed where a single transcript was in-frame (7T) and 4 transcripts were frameshifted as a result of slippage, most of the transcripts showed no signs of slippage. The original number of Ts in the knock-in *araC* gene is 8T.

The introduction of N protein decreases growth rate and cell fitness

Bacteriophage λ N protein has been reported to stabilise the RNA-DNA hybrid during transcription, and thus prevents slippage between the nascent transcript and the DNA template (Parks et al., 2014). To examine whether N protein would reduce cell fitness by preventing RNAP slippage, plasmids carrying the N gene (pNAS150) were transformed into REL607 and REL607::*araC*. The resulting strains were grown in Davis minimal supplemented with 1% arabinose (DMA10K) and the growth rates were monitored.

There was an observable difference in growth rate between wild-type REL607 and REL607 + pNAS150 (*P*-value = 7.34E-06, Kruskal-Wallis) (Figure 3.12). The doubling time increased from the calculated doubling time of wild-type REL607 of 61.0±0.4 min to 77.0±0.7 min and 96.5±1.9 min for REL607 + pNAS150 and REL607 + pNAS150 (induced with 1 mM IPTG), respectively ($\bar{X}$±SE, n=8). There was also an observable change in the growth rate between REL607::*araC* and REL607::*araC* + pNAS150 (*P*-value = 8.24E-06, Kruskal-Wallis) (Figure 3.12). The doubling time of REL607::*araC* + pNAS150 and REL607::*araC* + pNAS150 (induced with 1 mM IPTG) increased to 110.8±3.4 min and 130.0±1.7 min respectively, from the calculated doubling time of 90.3±1.0 min for REL607::*araC* ($\bar{X}$±SE, n=8). Overall, there was an increase in doubling time for the strains carrying pNAS150 compared to their cognate strains without pNAS150.

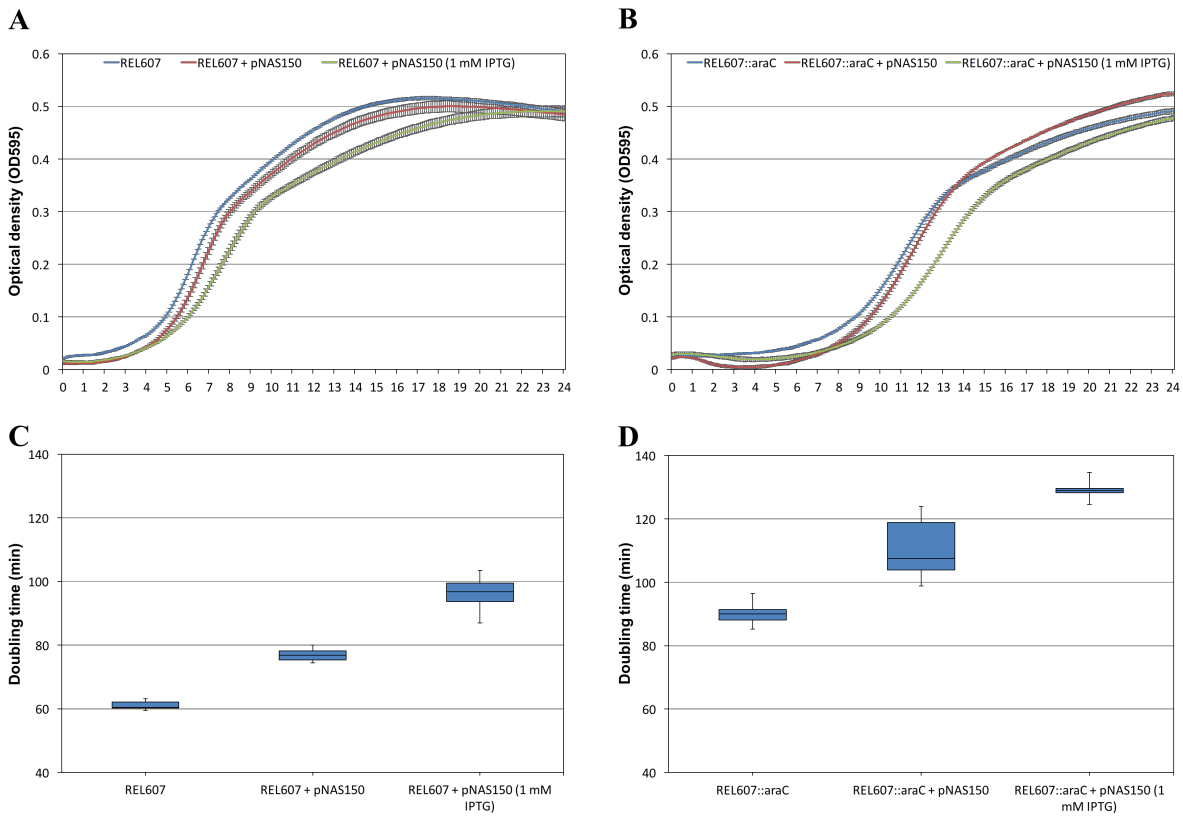

Figure 3.12 | The introduction of N protein resulted in an increase in doubling time and loss of fitness. Three REL607 and REL607::*araC,* and six REL607 + pNAS150 and REL607::*araC* + pNAS150 clones were grown in DMA10K media at 37°C with shaking. The OD595 points are shown as $\bar{X}$±SD, n=8. **A)** The OD595 of REL607 and REL607 + pNAS150 (with and without IPTG) were monitored for 24 hours. **B)** The OD595 of REL607::*araC* and REL607::*araC* + pNAS150 (with and without IPTG) were monitored for 24 hours. **C)** The doubling time of REL607, REL607 + pNAS150, and REL607 + pNAS150 (induced with IPTG) were 61.0±0.4 min, 77.0±0.7 min and 96.5±1.9 min respectively. **D.** The doubling time of REL607::*araC*, REL607::*araC* + pNAS150, and REL607::*araC* + pNAS150 (induced with IPTG) were 90.3±1.0 min, 110.8±3.4 min and 130.0±1.7 min respectively. Doubling times are shown as $\bar{X}$±SE, n=8.

Discussion

Studies on free-living bacterial genomes have shown that long polyA/T tracts are under-represented in bacterial genomes (Ackermann and Chao, 2006; Baranov et al., 2005; Orsi et al., 2010; Sharma et al., 2011). These tracts are highly selected against as they offer no inherent advantage to the cell. In some circumstances, these tracts might even be deleterious, when a gene that codes for a single protein acquires this tract as apposed to a gene that encodes multiple proteins. Remarkably, these polyA/T tracts were observed in 5-50% of genes in *Buchnera* (Tamas et al., 2008). Here, we asked whether these inherently disadvantageous polyA/T tracts could evolve in *E. coli* under conditions favouring drift. We evolved 10 genetically independent lines of *E. coli* with a mutator phenotype by streaking a starting plate (ancestor), randomly picking a single colony and streaking it onto a fresh LB plate (founder). This pick-streak-incubate process was repeated 100 times generating 10 independent lineages of *E. coli* populations evolving under intense genetic drift, allowing slightly deleterious mutations to accumulate in the genome (Felsenstein, 1974; Muller, 1964). The genomes of these bottlenecked and the non-bottlenecked control lineages were sequenced after 100 passages and we observed the emergence of 22 frameshifted polyA/T tracts (Figure 3.2 and Table 3.3).

These frameshifted polyA/T tracts require slippage-type editing to produce functional proteins from an otherwise non-functional gene, and upon further inspection, these genes (Table 3.3) were categorised as non-essential genes in the profiling of *E. coli* chromosome PEC (Yamazaki et al., 2008) and KEIO (Baba et al., 2006) databases, suggesting that these mutations may have arisen as a result of drift. Under conditions where selection is high, these polyA/T tracts would be eliminated from highly-expressed and conserved protein coding sequences as these tracts would result in erroneous transcripts and ultimately aberrant

proteins. This is consistent with the neutral evolution of pseudogenes (genes that resemble known genes but cannot produce functional proteins) where the rate of nucleotide substitution is higher in non-essential genes compared to essential genes, as the latter would be subjected to higher purifying selection (Li et al., 1981; Ohta, 1973). The emergence of these frameshifted polyA/T tracts in our evolution experiment is consistent with the neutral evolution theory where slippage-prone tracts may readily emerge and be fixed under conditions of drift (Covello and Gray, 1993; Stoltzfus, 1999).

A plausible origin for the evolution of RNA editing is that the compensating mechanism was already present under neutral conditions before a need for any corrective function was required through a process of pre-suppression (Gray, 2012, 2013; Gray et al., 2010). The idea behind pre-suppression is that a ratchet-like affinity increases in dependence, thus becoming interdependent over time (Gray et al., 2010) (discussed in Chapter 1). This model also suggests that an increase in dependency is inevitable if there are more ways for dependence to increase than decrease, and the two components eventually 'ratchet' toward greater dependency. These ratchet-like affinities may prevent the event of neutral fixed complexity to be neutrally unfixed through random reversion (Gray et al., 2010). This idea of pre-suppression is consistent with Covello and Gray's (1993) 3-step neutral model for the evolution of RNA editing (Box 3.1). I will discuss and elaborate on the 3-steps in regards to the emergence of slippage-type editing.

Box 3.1 | 3-step model for the evolution of RNA editing
Step 1: The appearance of the RNA editing machinery before there is need for editing
Step 2: Mutation at editable nucleotide positions and fixation by drift
Step 3: Maintenance of RNA editing by selection

In the first stage of the neutral model for the evolution of RNA editing, the editing activity pre-exists through the process of genetic drift, however, there are no substrates for the enzyme to act upon. Studies have shown that *E. coli* RNA polymerase is slippage-prone just like its close relatives, *Buchnera* and *Blochmannia* (Tamas et al., 2008; Wagner et al., 1990; Wernegreen et al., 2010), however, upon whole genome sequencing, we observed an under-representation of slippage-prone tracts in *E. coli* REL606 (Figure 3.2) compared to *Buchnera* where between 5 to 50% of coding sequences were riddled with slippage-prone tracts (Tamas et., 2008). It is common to assume that there will be a strong selection for improved function or fidelity of RNA polymerase because RNA polymerase slippage does not provide any selective advantage, and in most circumstances, slippage actually reduces gene expression efficiency as observed in *Buchnera* and *Blochmannia* (Tamas et al., 2008). In the low abundance of slippage-prone polyA/T tracts in *E. coli* REL606, there is little to no selection against the RNA polymerases' slippage-prone characteristic, enabling it to persist despite having no inherent benefit to the cells. This slippage-prone characteristic was observed in *E. coli* REL606 with the production of *araC* transcripts with varying lengths (Figure 3.11). The presence of a slippery RNA polymerase and the absence of slippage-prone tracts in *E. coli* RNA polymerase are consistent with the neutral model of RNA editing, where the enzyme is present before there is a substrate for the enzyme to act upon.

PolyA/T tracts are under-represented in the genomes of *E. coli* but upon inspection of the *E. coli* REL606 genome, we discovered thousands of sites in the genome where a single base change would produce a long polyA/T tract (Supplementary Table 3.2). Mutations that change a single base to an A/T, or the addition or removal of A/Ts, produce long polyA/T tracts and such mutations can be fixed in the genome by chance and we saw this occurring in 22 genes under our experimental conditions (Figure 3.2). The emergence and subsequent

fixation of these frameshifted slippage-prone polyA/T tracts is consistent with Muller's ratchet, where under conditions of high genetic drift and inefficient selection, slightly deleterious mutations such as frameshift polyA/T may be fixed in the genome (Felsenstein, 1974; Muller, 1964).

Upon fixation of these frameshifted slippage-prone polyA/T tracts, a situation now arises where RNA polymerase slippage is important for producing functional proteins. Slippage-type editing thus becomes an indispensable part of the genetic information pathway, correcting the reading frames of frameshifted pseudogenes at the RNA level, and is subsequently maintained by selection. Although we have not explicitly shown the third step of the model, we have performed some experiments to test this step (see Appendix). This model of RNA editing evolution is an example of a gratuitous complex process that fixes novel mechanisms without conferring any inherent advantage or increase of functionality (Covello and Gray, 1993; Gray et al., 2010; Lukeš et al., 2011; Stoltzfus, 1999). Our results on the emergence of slippage-prone tracts that require editing-type processes to produce functional proteins are consistent with Covello and Gray's (1993) model of the evolution of RNA editing.

We subjected *E. coli* populations to single-colony bottlenecks and we observed the emergence of slippage-prone polyA/T tracts (Figure 3.2). We observed the addition of a thymine residue to the existing poly 7T in *araC* that resulted in a frameshift mutation (this mutation is the only mutation observed in *araC*). We then asked whether this frameshift mutation affected cell fitness. To test the effect of frameshifted *araC* gene on cell fitness, we introduced this frameshifted *araC* gene, and knocked out *araC* from wild-type REL606 and REL607, and measured the growth rates. There was no observable difference in growth rates

when the strains were grown in rich LB medium (Figure 3.8), suggesting that *araC* is non-essential under this particular experimental condition. To reinforce this, the PEC database (Yamazaki et al., 2008) and KEIO knockout collection (Baba et al., 2006) have both listed *araC* as a non-essential gene in rich media.

Early studies have shown that glucose is the preferred carbon source of *E. coli* that supports the fastest growth of this bacterium (Walker et al., 1934), while more recent studies have shown that the hierarchical order of sugar preference for *E. coli* is glucose, followed by arabinose, sorbitol, and galactose (Perez-Alfaro et al., 2014). Therefore, in the absence of glucose in the media, and with arabinose as the sole carbon source, we would expect that the deletion or disruption the *araC* gene would be detrimental. To test the effect of *araC* under conditions where arabinose is the rate-limiting factor, we monitored the growth of the REL606 and REL607 (frameshifted *araC* knock-in and *araC* knock-out) strains in minimal media supplemented with arabinose (DMA10K). No growth was observed for REL606 and its derived strains (Figure 3.9) in DMA10K that consist of arabinose as the sole carbon source, and this result is expected as REL606 is an Ara^- strain incapable of metabolising arabinose (Jeong et al., 2009; Studier et al., 2009).

We subsequently measured the fitness of REL607 (and its derived strains), capable of metabolising arabinose in DMA10K, to assess the effect of *araC* on fitness. No growth was observed for REL607Δ*araC* (Figure 3.9) and RT-PCR results demonstrated that no *araC* transcripts were expressed (Figure 3.7), rendering the strain unable to metabolise arabinose from the media. Studies have reported that AraC regulates the *ara* operon that is required for the uptake and catabolism of L-arabinose (Bustos and Schleif, 1993; Schleif, 2000, 2010). The AraC protein functions as a homodimer (Schleif, 2003, 2010) and our homology

modelling of the frameshifted AraC suggested that our modelled AraC cannot form homodimers as the protein is missing the antiparallel coiled-coil region that forms the dimerisation interface (Figure 3.5) (Soisson et al., 1997). Interestingly, growth was observed in DMA10K along with an increase in doubling time in REL607::*araC* (61±0.3 min) compared to wild-type REL607 (90±1.0 min) (Figure 3.9). These results suggest that slippage-type editing rescued *araC* gene expression by the generation of a proportion of in-frame *araC* transcripts that would subsequently be translated to full-length AraC.

We also observed a longer lag phase for REL607::*araC* where the lag phase lasted for approximately 3 hours compared to wild-type where the lag phase was approximately an hour long (Figure 3.9). The lag phase of bacterial growth is poorly understood due to the low densities of cells and low metabolic activity. However, Madar and colleagues overcame these limitations by developing an assay based on imaging flow cytometry of fluorescent reporter cells that permits the measurement of individual cell size, cell number and promoter activity (Madar et al., 2013). This study demonstrated that gene expression during early lag phase prioritises carbon source utilisation enzymes over biomass accumulation (Madar et al., 2013; Schultz and Kishony, 2013). Madar and colleagues (2013) discovered that a cell does not accumulate biomass as early as possible, but instead produces enzymes that function in uptake and catabolism of carbon sources. The long lag phase that we observed in our study could be a result of a low proportion of functional AraC being produced from the frameshifted *araC* as a result of RNA polymerase slippage. With the production of a small subset of functional AraC, we would expect to see an increase in levels of expression of *araC* to compensate for the production of erroneous transcripts enabling the production of functional AraC for the utilisation of arabinose. It is plausible that the cells could be expending more resources on utilising arabinose as efficiently as possible through increased

araC expression, at the expense of cell biomass (Goldsmith and Tawfik, 2009). Thus, the observed long lag phase (low optical density) in our REL607::*araC* strain, would be consistent with Madar and colleagues' hypothesis of prioritising the utilisation of carbon sources over cell biomass. In order to test this, we could assess the *araC* expression levels by performing qRT-PCR and ask whether there will be a change in gene expression levels to overcome the production of erroneous *araC* mRNAs.

We have also tested whether the growth observed in arabinose only media could be attributed to remnants of the LB media used to grow our overnight cultures. This was tested by growing our overnight cultures in LB and diluting 1:100 in Davis minimal without any carbon source (DM0) and monitoring the optical density on the plate reader. We would expect to see growth in strains incapable of utilising arabinose if the glucose present in LB was contributing to the observed growth but instead, we have not seen a marked increase in growth (Figure 3.10). This result suggests that the small amount of LB remnant is not sufficient to support the growth of Ara$^-$ strains and the observed growth in arabinose only media is due to the ability to utilise arabinose present in the media.

We then examined whether slippage-type editing can rescue gene expression of the frameshifted *araC* by assessing the mRNA transcripts of REL607::*araC*. We have previously shown that AraC is not essential under conditions where other more favourable carbon sources are available (Figure 3.8). Therefore, to assess the impact of RNA polymerase slippage on *araC* expression, we extracted the RNA from cells grown in minimal media supplemented with arabinose only (DMA10K). We observed heterogeneity in the *araC* transcripts (Figure 3.11), and although the level of slippage was low, we observed some proportion of transcripts that were frameshifted and one transcript was in the correct frame.

These results suggested that RNA polymerase infidelity resulted in a heterogeneous population of mRNAs as previously observed *in vitro* in *E. coli* (Wagner et al., 1990) and *in-vivo* in endosymbionts (Tamas et al., 2008). The reduction in *araC* expression efficiency when grown in DMA10K may address the loss of fitness observed in REL607::*araC* compared to wild-type REL607. This is because a proportion of AraC protein produced is in the correct frame, and a proportion of them were frameshifted due to RNA polymerase slippage. Although we only observed a single *araC* transcript that was corrected, studies have shown that a single RNA transcript can be translated multiple times, where protein molecules are produced in bursts, with each burst originating from a stochastically transcribed single mRNA molecule (Yu et al., 2006). To test this phenomenon of bursts in gene expression in regards to *araC* expression, we could assess individual *araC* mRNAs within single cells by means of single-cell RT-PCR (Warren et al., 2006) termed digital RT-PCR, which is a variation on digital PCR (Vogelstein and Kinzler, 1999).

We previously assessed the effect of frameshifted *araC* on cell fitness by monitoring the growth rates in arabinose only media, where we observed growth in cells bearing frameshifted *araC* despite a loss of fitness (Figure 3.9). Additionally, we observed heterogeneity of *araC* RNA transcripts (Figure 3.11) and we suggested that slippage-type editing rescued *araC* expression by correcting the reading frame. Recent studies on bacteriophage λ N protein suggested that N reduces the frequency of RNA polymerase slippage by stabilising the upstream end of the RNA-DNA hybrid (Parks et al., 2014). Therefore, upon the introduction of N, we would expect to see a further decline in fitness where slippage-type editing is required for the production of functional proteins as RNA polymerase slippage is inhibited by N. Here, we examined the effect of N protein on cell fitness by transforming the N-bearing pNAS150 plasmid into REL607::*araC*. We observed a

decrease in growth rate upon the introduction of pNAS150, and a further decrease in growth rate was observed upon the induction of N protein expression (Figure 3.12). The loss of fitness in the presence of N protein could be explained by the inhibition of RNA polymerase slippage on the frameshifted *araC* gene however, we also observed a loss of fitness in the wild-type REL607 that bears an in-frame *araC* gene. The loss of fitness could, therefore, be explained by the inhibition of expression of genes that requires slippage-type editing for the production of functional proteins (apart from the frameshifted *araC*). Although sequences that are slippage-prone occur infrequently, slippage-type editing has been shown to have regulatory roles in *E. coli pyrBI*, *codBA* and *upp* operons (Liu et al., 1994; Qi and Turnbough Jr, 1995; Tu and Turnbough, 1997) (see Chapter 1), and the inhibition of RNA polymerase slippage would, therefore, be deleterious in these cases.

Studies showed that N protein binds to a specific RNA hairpin structure within a cis-acting DNA sequence called *nut* for efficient suppression of termination signals (Das, 1993; Mason et al., 1992; Roberts et al., 2008). Studies have demonstrated that N alone is sufficient to cause partially processive antitermination when over-expressed *in vivo* and added in excess of elongation complexes *in vitro* (Nudler and Gusarov, 2003; Rees et al., 1996). The over-expression of N *in vivo* inhibits growth, although the reason behind this deleterious effect was not studied in detail (Court, NIH, personal communication). The presence of a *nut* binding site within the *E. coli* chromosome and the overall deleterious effects of over-expression of N protein would, therefore, prevent accurate assessment of the effect of N protein in preventing transcriptional slippage in our current experimental setup. Consequently, to directly test the effect of N protein on cell fitness *in vivo*, a *nut* site could be placed upstream of *araC* gene and the high expression vector (pNAS150, derived from pUC9) could be substituted with a lower copy number plasmid with more stringent regulation of N protein expression.

To summarise, although slippage-prone tracts are highly selected against in free-living *E. coli* populations, we have shown that these tracts may readily evolve under lab conditions favouring drift. Furthermore, we demonstrated that RNA polymerase slippage resulted in transcript heterogeneity and a reduction in cell fitness under certain conditions. Nevertheless, in this chapter we did not examine the impact of transcriptional slippage on protein production, and we go on to further examine this in Chapter 4.

Supplementary

Table 3.1 | Functional categories for protein-coding genes containing polyA/T tracts of 9 nucleotides or more

Functional categories		Day 50		Day 100
		Poly 9A/T	Poly 10A/T	Poly 9A/T
Information storage	J Translation, ribosome structure 14 89 40 93	**0**		**0**
	A RNA processing/modification 0 1 0 0	**0**		**0**
	K Transcription 3 16 5 10	**1**		**0**
	L Replication, recombination 9 33 23 29	**0**		**1**
Cellular processes	D Cell cycle control 1 8 5 6	0		0
	V Defense mechanisms 1 3 2 2	0		0
	T Signal transduction 0 4 3 3	0		1
	M Cell wall/membrane 6 23 2 3	2		2
	N Cell motility 1 23 9 12	0		1
	U Intracellular trafficking 4 24 4 8	0		0
	O Posttranslational modification 4 39 13 19	2		2
Metabolism	C Energy production 5 41 14 28	0		**0**
	G Carbohydrate transport/metabolism 2 22 8 16	0		0
	E Amino acid transport/metabolism 5 55 15 25	0		1
	F Nucleotide transport/metabolism 2 23 3 5	0		0
	H Coenzyme transport/metabolism 1 30 4 5	1		2
	I Lipid transport/metabolism 1 12 6 7	0		0
	P Inorganic ion transport/metabolism 1 15 3 5	0		1
	Q Secondary metabolites 0 2 1 1	0		0
Poorly characterized	R General function prediction only 8 29 11 12	1	1	2
	S Function unknown 3 14 7 12	1		1
TOTAL		8	1	14

Table 3.2 | Number of potential slippage-prone sites* in the genome of *E. coli* B strain REL606

Number of bases (A/T)	Adenine tract with one mismatch	Thymine tract with one mismatch
10	490	553
11	129	174
12	28	62
13	4	17
14	0	3

*A single change, either a deletion or substitution to A/T base to the runs of A/Ts will produce a polyA/T tract

Appendix

We have shown that slippage-prone polyA/T tracts can readily emerge under conditions of strong drift. These results were consistent with the 3-step model for the evolution of RNA editing, but we have not provided concrete evidence for the persistence of this editing-type process in *E. coli*. The final step of the model states that slippage-type editing is indispensable for gene expression and is maintained by selection following the fixation of mutations. To test this, we carried out an evolution experiment, where the bottlenecked lineage, BN50.1 (day 50 of the lineage BN100.1 from the evolution experiment), was subjected to a bottleneck relief regime (results not shown). This regime consisted of serial transfer of large populations, of *E. coli* (similar to the method used for the transfer of the non-bottlenecked lineages), creating a condition where selection may act more efficiently. Under bottleneck relief conditions, we would expect to see selection operating more efficiently in removing deleterious mutations. More importantly, if slippage-type editing is in fact integral to gene expression, we would expect to see the persistence of these observed frameshifted polyA/T tract in the genome as functional protein may be produced from these pseudo-pseudogenes by means of slippage-type editing.

Using BN50.1 (subjected to 50 days of bottlenecking) as the initiating strain, we subjected the line to 120 days of the bottleneck relief regime and sequenced the whole-genome at days 117, 118 and 119 (termed BR117, BR118, and BR119). Whole-genome analysis revealed that these slippage-prone tracts persisted in the genome of BR117 however, subsequent Delta-bitscore (DBS) analysis of BR117 revealed that most of the highly deleterious mutations reverted back to the wild-type state, suggesting that we were dealing with a mixed population of bacteria. Furthermore, BR118 and BR119 sequences were both wild-type REL606, suggesting that our bottleneck relief lines were contaminated with wild-type

REL606 in day 117 and by day 118, the much fitter wild-type REL606 have taken over the entire population. Genome sequencing and mapping suggested that a proportion of BR117 genomes were descendants of BN50.1 as we observed approximately 3,800 mutations that were also observed in BN50.1 including the slippage-prone tracts. By performing whole transcriptome sequencing (RNA-seq) we can assess whether slippage-type editing is, in fact, proceeding along these slippage-prone tracts by checking for heterogeneity of each read. RNA-seq will not only allow us to examine transcript heterogeneity, but it can also provide details on gene expression levels. As RNA polymerase slippage produces a proportion of transcripts of varying lengths, we would expect to see an increase in the level of gene expression of genes bearing slippage-prone polyA/T tracts to increase the likelihood of the production of full-length transcripts. Further to this, to test whether full-length proteins are produced, we can perform western blot and mass spectrophotometry.

References

Ackermann, M., and Chao, L. (2006). DNA sequences shaped by selection for stability. PLoS Genet. *2*, e22.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. *2*, 2006.0008.

Baranov, P.V., Hammer, A.W., Zhou, J., Gesteland, R.F., and Atkins, J.F. (2005). Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. Genome Biol. *6*, R25.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. *42*, W252-258.

Bustos, S.A., and Schleif, R.F. (1993). Functional domains of the AraC protein. Proc. Natl. Acad. Sci. U. S. A. *90*, 5638–5642.

Chamberlin, M., and Berg, P. (1962). Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from *Escherichia coli*. Proc. Natl. Acad. Sci. *48*, 81–94.

Covello, P.S., and Gray, M.W. (1993). On the evolution of RNA editing. Trends Genet. TIG *9*, 265–268.

Das, A. (1993). Control of transcription termination by RNA-binding proteins. Annu. Rev. Biochem. *62*, 893–930.

Edwards, M.D., Black, S., Rasmussen, T., Rasmussen, A., Stokes, N.R., Stephen, T.-L., Miller, S., and Booth, I.R. (2012). Characterization of three novel mechanosensitive channel activities in *Escherichia coli.* Channels Austin Tex *6*, 272–281.

Ezraty, B., Aussel, L., and Barras, F. (2005). Methionine sulfoxide reductases in prokaryotes. Biochim. Biophys. Acta *1703*, 221–229.

Fehér, T., Karcagi, I., Gyorfy, Z., Umenhoffer, K., Csörgo, B., and Pósfai, G. (2008). Scarless engineering of the *Escherichia coli* genome. Methods Mol. Biol. Clifton NJ *416*, 251–259.

Felsenstein, J. (1974). The evolutionary advantage of recombination. Genetics *78*, 737–756.

Franke, S., Grass, G., Rensing, C., and Nies, D.H. (2003). Molecular analysis of the copper-transporting efflux system CusCFBA of *Escherichia coli*. J. Bacteriol. *185*, 3804–3812.

Goldsmith, M., and Tawfik, D.S. (2009). Potential role of phenotypic mutations in the evolution of protein expression and stability. Proc. Natl. Acad. Sci. *106*, 6197–6202.

Gray, M.W. (2012). Evolutionary origin of RNA editing. Biochemistry (Mosc.) *51*, 5235–5242.

Gray, M.W. (2013). RNA editing: Evolutionary implications. In eLS, John Wiley & Sons, Ltd, ed. (Chichester, UK: John Wiley & Sons, Ltd).

Gray, M.W., Lukeš, J., Archibald, J.M., Keeling, P.J., and Doolittle, W.F. (2010). Irremediable complexity? Science *330*, 920–921.

Jeong, H., Barbe, V., Lee, C.H., Vallenet, D., Yu, D.S., Choi, S.-H., Couloux, A., Lee, S.-W., Yoon, S.H., Cattolico, L., et al. (2009). Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). J. Mol. Biol. *394*, 644–652.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinforma. Oxf. Engl. *28*, 1647–1649.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Li, W.H., Gojobori, T., and Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution. Nature *292*, 237–239.

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res. Notes *5*, 337.

Lindgreen, S., Krogh, A., and Pedersen, J.S. (2014). SNPest: a probabilistic graphical model for estimating genotypes. BMC Res. Notes *7*, 698.

Liu, C., Heath, L.S., and Turnbough, C.L. (1994). Regulation of *pyrBI* operon expression in *Escherichia coli* by UTP-sensitive reiterative RNA synthesis during transcriptional initiation. Genes Dev. *8*, 2904–2912.

Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F., and Gray, M.W. (2011). How a neutral evolutionary ratchet can build cellular complexity. IUBMB Life *63*, 528–537.

Madar, D., Dekel, E., Bren, A., Zimmer, A., Porat, Z., and Alon, U. (2013). Promoter activity dynamics in the lag phase of *Escherichia coli.* BMC Syst. Biol. *7*, 136.

Maddamsetti, R., Lenski, R.E., and Barrick, J.E. (2015). Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with *Escherichia coli*. Genetics *200*, 619–631.

Mason, S.W., Li, J., and Greenblatt, J. (1992). Host factor requirements for processive antitermination of transcription and suppression of pausing by the N protein of bacteriophage lambda. J. Biol. Chem. *267*, 19418–19426.

Metcalf, W.W., and Wanner, B.L. (1991). Involvement of the *Escherichia coli phn (psiD)* gene cluster in assimilation of phosphorus in the form of phosphonates, phosphite, Pi esters, and Pi. J. Bacteriol. *173*, 587–600.

Muller, H.J. (1964). The relation of recombination to mutational advance. Mutat. Res. *106*, 2–9.

Nudler, E., and Gusarov, I. (2003). Analysis of the intrinsic transcription termination mechanism and its control. Methods Enzymol. *371*, 369–382.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature *246*, 96–98.

Orsi, R.H., Bowen, B.M., and Wiedmann, M. (2010). Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. BMC Genomics *11*, 102.

Parks, A.R., Court, C., Lubkowska, L., Jin, D.J., Kashlev, M., and Court, D.L. (2014). Bacteriophage λ N protein inhibits transcription slippage by *Escherichia coli* RNA polymerase. Nucleic Acids Res. *42*, 5823–5829.

Perez-Alfaro, R.S., Santillan, M., Galan-Vasquez, E., and Martinez-Antonio, A. (2014). Regulatory switches for hierarchical use of carbon sources in *E. coli*. Netw. Biol.

Qi, F., and Turnbough Jr, C.L. (1995). Regulation of *codBA* operon expression in *Escherichia coli* by UTP-dependent reiterative transcription and UTP-sensitive transcriptional start site switching. J. Mol. Biol. *254*, 552–565.

Rees, W.A., Weitzel, S.E., Yager, T.D., Das, A., and von Hippel, P.H. (1996). Bacteriophage lambda N protein alone can induce transcription antitermination in vitro. Proc. Natl. Acad. Sci. *93*, 342–346.

Rhoads, D.B., Laimins, L., and Epstein, W. (1978). Functional organization of the *kdp* genes of *Escherichia coli* K-12. J. Bacteriol. *135*, 445–452.

Roberts, J.W., Shankar, S., and Filter, J.J. (2008). RNA Polymerase Elongation Factors. Annu. Rev. Microbiol. *62*, 211.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, N. Y.: Cold Spring Harbor Laboratory Pr).

Schleif, R. (2000). Regulation of the l-arabinose operon of *Escherichia coli*. Trends Genet. *16*, 559–565.

Schleif, R. (2003). AraC protein: a love-hate relationship. BioEssays News Rev. Mol. Cell. Dev. Biol. *25*, 274–282.

Schleif, R. (2010). AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. FEMS Microbiol. Rev. *34*, 779–796.

Schmitt, M.E., Brown, T.A., and Trumpower, B.L. (1990). A rapid and simple method for preparation of RNA from *Saccharomyces cerevisiae*. Nucleic Acids Res. *18*, 3091–3092.

Schrödinger, L. (2015). The PyMOL molecular graphics systems (Schrödinger, LLC).

Schultz, D., and Kishony, R. (2013). Optimization and control in bacterial Lag phase. BMC Biol. *11*, 120.

Sharma, V., Firth, A.E., Antonov, I., Fayet, O., Atkins, J.F., Borodovsky, M., and Baranov, P.V. (2011). A pilot study of bacterial genes with disrupted ORFs reveals a surprising

profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. Mol. Biol. Evol. *28*, 3195–3211.

Soisson, S.M., MacDougall-Shackleton, B., Schleif, R., and Wolberger, C. (1997). Structural basis for ligand-regulated oligomerization of AraC. Science *276*, 421–425.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. J. Mol. Evol. *49*, 169–181.

Studier, F.W., Daegelen, P., Lenski, R.E., Maslov, S., and Kim, J.F. (2009). Understanding the differences between genome sequences of *Escherichia col*i B strains REL606 and BL21(DE3) and comparison of the *E. coli* B and K-12 genomes. J. Mol. Biol. *394*, 653–680.

Tamas, I., Wernegreen, J.J., Nystedt, B., Kauppinen, S.N., Darby, A.C., Gomez-Valero, L., Lundin, D., Poole, A.M., and Andersson, S.G.E. (2008). Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. Proc. Natl. Acad. Sci. *105*, 14934–14939.

Taylor, F.R., Grogan, D.W., and Cronan, J.E. (1981). Cyclopropane fatty acid synthase from Escherichia coli. Methods Enzymol. *71 Pt C*, 133–139.

Tu, A.H., and Turnbough, C.L. (1997). Regulation of *upp* expression in *Escherichia coli* by UTP-sensitive selection of transcriptional start sites coupled with UTP-dependent reiterative transcription. J. Bacteriol. *179*, 6665–6673.

Vogelstein, B., and Kinzler, K.W. (1999). Digital PCR. Proc. Natl. Acad. Sci. *96*, 9236–9241.

Wagner, L.A., Weiss, R.B., Driscoll, R., Dunn, D.S., and Gesteland, R.F. (1990). Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli.* Nucleic Acids Res. *18*, 3529–3535.

Walderhaug, M.O., Polarek, J.W., Voelkner, P., Daniel, J.M., Hesse, J.E., Altendorf, K., and Epstein, W. (1992). KdpD and KdpE, proteins that control expression of the *kdpABC* operon, are members of the two-component sensor-effector class of regulators. J. Bacteriol. *174*, 2152–2159.

Walker, H.H., Winslow, C.E., and Mooney, M.G. (1934). Bacterial cell metabolism under anaerobic conditions. J. Gen. Physiol. *17*, 349–357.

Warren, L., Bryder, D., Weissman, I.L., and Quake, S.R. (2006). Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc. Natl. Acad. Sci. *103*, 17807–17812.

Watson, N., Dunyak, D.S., Rosey, E.L., Slonczewski, J.L., and Olson, E.R. (1992). Identification of elements involved in transcriptional regulation of the *Escherichia coli cad* operon by external pH. J. Bacteriol. *174*, 530–540.

Wernegreen, J.J., Kauppinen, S.N., and Degnan, P.H. (2010). Slip into something more functional: selection maintains ancient frameshifts in homopolymeric sequences. Mol. Biol. Evol. *27*, 833–839.

Yamazaki, Y., Niki, H., and Kato, J. (2008). Profiling of *Escherichia coli* chromosome database. Methods Mol. Biol. Clifton NJ *416*, 385–389.

Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X.S. (2006). Probing gene expression in live cells, one protein molecule at a time. Science *311*, 1600–1603.

CHAPTER 4
The impact of slippage-type editing on gene expression

Introduction

Errors may occur at all stages of genetic information transfer, from DNA to RNA through to protein. While errors in DNA replication may not directly dictate the final protein sequence, errors that arise during transcription and translation are more likely to disrupt downstream protein production. The error rate in DNA is significantly lower than the rates observed at the RNA and protein levels, at roughly 10^{-9} (Schaaper, 1993), compared to approximately 10^{-5} and 10^{-4} respectively for RNA and protein (Meyerovich et al., 2010; Ninio, 1991). Further to this, since a single mRNA transcript can be translated multiple times (Ozbudak et al., 2002; Raj and van Oudenaarden, 2008), RNA errors can become exponentially amplified, burdening the cell with misfolded, aberrant, and toxic proteins, under some circumstances. Despite this, studies on transcriptional error have received less attention than errors during replication, likely due to the fact that errors in transcription are considered transient and non-heritable (Pál and Hurst, 2000).

Homopolymeric tracts, particularly long polyA/T tracts, are thermodynamically unstable and are prone to enzymatic slippage (Koch, 2004; Viguera et al., 2001; Wagner et al., 1990). RNA polymerase slippage along such homopolymeric tracts results in a heterogeneous population of mRNAs that differ from their encoding DNA template (Tamas et al., 2008; Wagner et al., 1990). RNA polymerase slippage yields transcripts containing or lacking one

or more additional base(s) corresponding to the slippage-prone polyA/T sequence, thus introducing indels into RNA transcripts. Indels in polyA/T tracts often have deleterious effects because proteins that are translated from these erroneous mRNAs are likely to exhibit partial function, changes in function, loss of function, or in rare cases, a gain of function (Baranov et al., 2005). Therefore, homopolymeric A/T tracts are highly selected against in most free-living prokaryotes including *E. coli* (Baranov et al., 2005; Orsi et al., 2010). Although, RNA polymerase slippage has been reported in eukaryotes (Linton et al., 1997), bacteriophages (Macdonald et al., 1993), viruses (Ratinier et al., 2008) and prokaryotes (Tamas et al., 2008; Wagner et al., 1990). Although slippage-type editing appears to offer no inherent advantage to the cell, multiples studies have demonstrated that slippage-type editing is utilised for regulation of gene expression (Han and Turnbough, 2014; Turnbough, 2011), production of multiple proteins from a single gene (Hausmann et al., 1999; Penno et al., 2006), and the generation of fusion proteins (Schurig et al., 1995).

Homopolymeric tracts are also prone to ribosomal frameshifting, the alternate translation of mRNA sequences directed by cis-acting elements within the mRNA (Dinman, 2012a). These cis-acting elements generally include a slippery shift site of X.XXZ.ZZN (where XXX and ZZZ are triplets of identical bases, and N is any nucleotide), a short spacer sequence, and stimulatory motifs such as shine-Dalgarno-like sequences (SD-like) and pseudoknots (Dinman, 2012b, 2012a). Under certain circumstances, SD-like sequences upstream of the shift site help to position the ribosome at the shift site (Larsen et al., 1994), and the presence of these stimulatory motifs dramatically increases frameshifting rates (Atkins et al., 2016; Gurvich et al., 2003). Interestingly, the slippery motif (when it is A.AAA.AAN or T.TTT.TTN), which is a mandatory element in ribosomal frameshifting (Sharma et al., 2014), is also prone to RNA polymerase slippage.

In Chapter 3, we observed the emergence slippage-prone homopolymeric tracts in *E. coli* populations subjected to serial bottlenecks and demonstrated that these tracts are generally detrimental to cell fitness under certain conditions. In addition to this, RNA sequencing revealed that *E. coli* RNA polymerase is indeed slipping on these tracts, producing a heterogeneous population of mRNAs. Although we anticipated that this would lead to the production of a proportion of full-length protein along with a subset of potentially deleterious and truncated polypeptides, we have not yet assessed the effect of slippage on protein production. To our knowledge, there have been no reports on the impact of slippage-type editing on protein production.

In this study, we assessed the impact of RNA polymerase slippage on gene expression and protein production utilising GFP reporter systems. We performed RT-PCR to investigate the impact of slippage-type editing at the RNA level. Consequently, through western blotting of total protein, we demonstrate that RNA polymerase slippage rescues the function of genes with frameshift mutations by producing a proportion of full-length proteins. We also assessed whether the production of these GFP proteins of varying lengths are products of ribosomal frameshifting as opposed to slippage-type editing by analysing RNA secondary structure.

Methods

Strains and media

All chemicals were purchased from Sigma-Aldrich Co. unless otherwise specified. All oligonucleotides were synthesised by Integrated DNA Technologies. *E. coli* B strain REL606 was obtained from T. Cooper (University of Houston, Texas). REL606 and REL606-derived strains were grown at 37°C in Davis Minimal Broth (Difco) supplemented with 2 mg/L thiamine and 2000 mg/L dextrose (DM2000). For solid media, bacteriological agar (Oxoid) was added to a final concentration of 1.5% w/v. All experiments were conducted in the presence of antibiotics at the following concentrations: streptomycin, 100µg/mL and ampicillin, 100µg/mL (Peptides International). *Saccharomyces cerevisiae* S288c was obtained from Austen Ganley (University of Auckland), and grown at 30°C in yeast peptone dextrose (YPD) (Difco).

Construction of a slippage-prone GFP reporter system in E. coli

The GFP reporter systems consist of a ribosome-binding site (RBS), followed by a start site, a 6 X HisTag and a modified GFP gene (Figure 4.1). The slippage-prone polyA tract was designed by replacing the G bases with A bases in the AG.AAG.AAG.AA sequence at amino acid positions 157 to 160 of GFP. These changes to the codons do not change the amino acid sequence, and it has been established that the insertion of short sequences into this region does not affect GFP fluorescence (Paramban et al., 2004). The GFP genes with slippage-prone polyA tracts were synthesised and cloned into pUC57 by GenScript (USA). The GFP genes were then amplified using Phusion® High-fidelity DNA Polymerase (Thermo Fisher), with GFP_F1 and GFP_R1 primers. The amplified PCR products were then cloned into pGEM®-T Easy Vector (Promega) and the plasmids were transformed into *E. coli* REL606 using calcium chloride and heatshock (Sambrook *et al.* 1989). Subsequently, the plasmids

were screened with M13F and GFP_F1 primers, and M13R and GFP_R1 primer pairs to check for directionality.

Primers used for cloning:

M13F - 5′ – GTAAAACGACGGCCAGT – 3′

M13R - 5′ – AGCGGATAACAATTTCACACAGGA – 3′

GFP_F1 - 5′ – AGGAGGACAGCAATGCATCA – 3′

GFP_R1 - 5′ – TTATTTGTATAGTTCATCCATGCCATG – 3′

The plasmids with GFP downstream of the *lac* promoter were purified using the PureLink® Quick Plasmid Miniprep Kit (Invitrogen) and sequenced bi-directionally with M13 primers (Macrogen). Plasmids with the correct orientation were then transformed into *E. coli* REL606 competent cells using calcium chloride and heat shock. The reporter systems generated for this experiment are listed in Figure 4.1 and will be referred to from this point forward as No Slip, Slip and Frameshift.

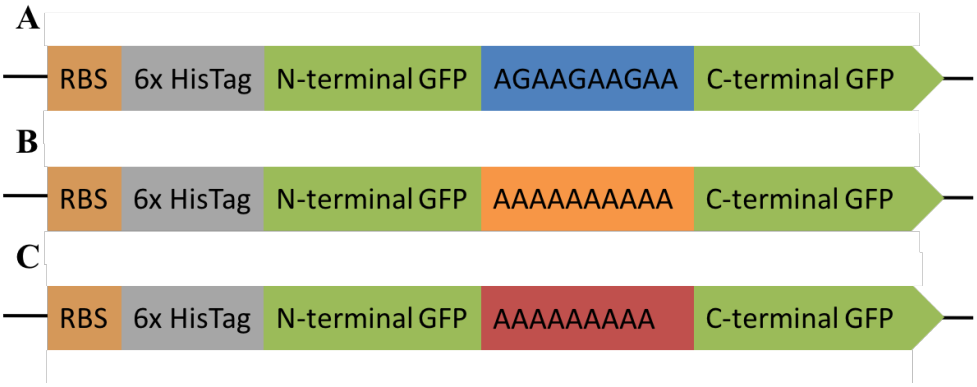

Figure 4.1 | The GFP reporter systems with varying polyA tracts. A) No slip construct, a GFP reporter consisting of in-frame non-slippage prone tract of $(AAG)_3A$. **B)** Slip construct, consisting of an in-frame slippage-prone tract of 10A. **C)** Frameshift construct, consisting of a frameshifted tract of 9A.

Site-directed mutagenesis

The Frameshift GFP reporter (Figure 4.1) was generated from the Slip GFP reporter using site-directed mutagenesis. The procedure for site-directed mutagenesis utilises a supercoiled double-stranded DNA (dsDNA) vector with an insert of interest and a pair of oligonucleotide primers containing the desired mutation (Figure 4.2). The oligonucleotide primers were designed to flank the poly 10A tract with a deletion of an adenine base (indicated by an asterisk) to generate a poly 9A tract. The parental plasmid pGEM::Slip was extracted from REL606, a dam^+ strain (Dam methylase positive) and amplified with the GFP_PolyA_SMut_F and GFP_PolyA_SMut_R primer pair. The PCR products were then subjected to DpnI endonuclease (target sequence: 5′-Gm6ATC-3′) digestion, where the parental methylated DNA was digested while the mutated plasmid containing staggered nicks remains intact. The nicked pGEM::Frameshift containing the desired mutation was then transformed into *E. coli* REL606 competent cells with calcium chloride and heat shock (Sambrook et al., 1989). The resulting plasmid was purified using the PureLink® Quick Plasmid Miniprep Kit (Invitrogen) and confirmed by sequencing with M13 primers (Macrogen).

Primers used for site-directed mutagenesis:

GFP_PolyA_SMut_F:

5′ – AAACTACGGCAAAAAAAAACGAACAGCC – 3′

GFP_PolyA_SMut_R:

5′ – CTTCAACCGAAACTACGGCAAAAAAAAACG – 3′

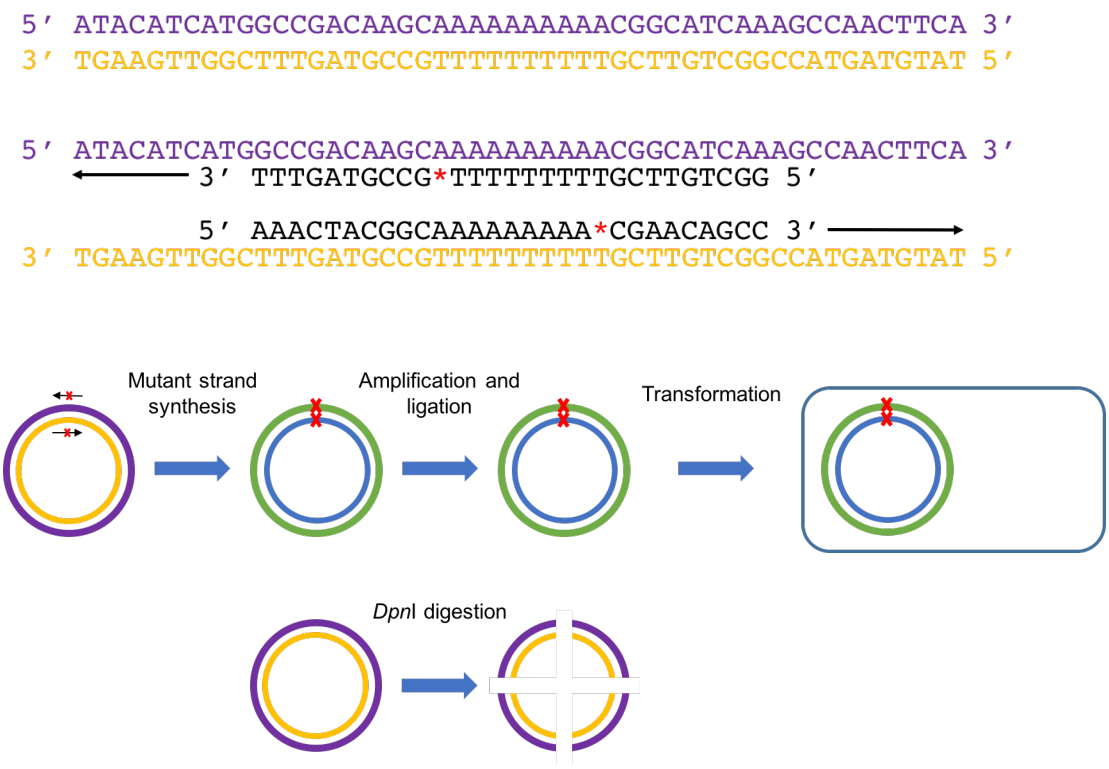

Figure 4.2 | Site-directed mutagenesis was utilised to generate a frameshifted 9A tract from the in-frame 10A tract. The parental plasmid, pGEM::Slip (pictured in purple and yellow), which carries a 10A tract within the GFP gene, was purified and amplified with a primer pair that consists of a 9A tract instead of 10A. The PCR product that consists of the mutated nicked plasmid (pictured green and blue) and parental plasmid were then subjected to DpnI digestion where the methylated parental plasmid is digested. The mutated plasmid with staggered nicks was then transformed into *E. coli* REL606.

Control experiments for enzyme slippage

To control for slippage of commercial reverse transcriptase, DNA polymerase and sequencing artefacts, we tested for slippage using *S. cerevisiae* S288c as a model system for no slippage. *S. cerevisiae* endogenous RNA polymerase has been shown to not be slippage-prone (Wagner et al., 1990), making it feasible to determine whether the exogenous enzymes used in this study are slippage-prone. To examine whether *S. cerevisiae* endogenous DNA

polymerase and commercial DNA polymerase are prone to slippage, we amplified the *S. cerevisiae* KRS1 gene (encodes Lysyl-tRNA synthetase), consisting of a poly 10A tract using Phusion® High-fidelity DNA Polymerase (Thermo Fisher) with KRS1_F and KRS1_R primers. The PCR products were subsequently purified using the Wizard SV Gel and PCR Clean- up System (Promega) and sequenced bi-directionally (Macrogen). If slippage is seen, it can be attributed to either experimental enzymes or sequencing artefacts. To assess whether *E. coli* endogenous DNA polymerase is slippage-prone, we extracted the plasmids using the PureLink® Quick Plasmid Miniprep Kit (Invitrogen), purified with the Wizard SV Gel and PCR Clean- up System (Promega) and sequenced the pGEM-T easy plasmids bearing the polyA tract (Macrogen).

To establish whether commercially available reverse transcriptase is slippage-prone, we prepared an overnight culture of *S. cerevisiae* in YPD at 30°C. The *S. cerevisiae* culture was then diluted 1:100 in 10 mL of fresh YPD and the culture was grown for approximately 6 hours at 30°C. Total RNA was extracted from the mid-log phase *S. cerevisiae* culture using a hot phenol method, previously described by Schmitt et al., (1990). The extracted total RNA was subsequently diluted to 300 ng/uL and then subjected to DNase I treatment using TURBO DNase I (Ambion), following the manufacturers' guidelines. The DNase I-treated total RNA was checked for genomic DNA contamination using Phusion® High-fidelity DNA Polymerase (Thermo Fisher) with KRS1_F and KRS1_R primers. The DNA-free total RNA was checked for mRNA quality prior to first-strand synthesis with gstA_fwd and gstA_rev primers, which amplify the house-keeping gene Glutathione S-transferase (GST), using SuperScript® One-Step RT-PCR System with Platinum® *Taq* DNA Polymerase kit (Invitrogen).

The DNA-free total RNA was then subjected to first-strand synthesis using SuperScript II (Invitrogen) with gene specific primers, KRS1_F and KRS1_R, following the manufacturers' guidelines. The first-strand cDNA was amplified with the KRS1_F and KRS1_R primer pair using Phusion® High-fidelity DNA Polymerase (Thermo Fisher). The amplified PCR products were sequenced bi-directionally (Macrogen) with the KRS1_F and KRS1_R primer pair. The amplified PCR products were also A-tailed with *Taq* Polymerase following the manufacturer's protocols (Bioline), cloned into pGEM-T easy (Promega) and transformed into *E. coli* DH5α using calcium chloride and heat shock (Sambrook et al., 1989). The transformants were screened with the M13 primer pair and KAPA2G Robust HotStart ReadyMix PCR Kit (KAPA Biosystems), and plasmids with the correct insert size were extracted and sequenced bi-directionally (Macrogen) with the M13 primer pair.

Primers used to detect KRS1 transcripts were:

KRS1_F:

5′ – ACGAAGCTACCGGGGAAATG – 3′

KRS1_R:

5′ – ACGTAACCCTCGACACCAAC – 3′

Primers used to detect *gstA* transcripts were:

gstA_fwd:

5′ – CTTTGCCGTTAACCCTAAGGG – 3′

gstA_rev:

5′ – GCTGCAATGTGCTCTAACCC – 3′

Total E. coli RNA extraction and RT-PCR

The REL606 GFP reporter systems were grown overnight in DM2000 at 37°C. The bacterial cultures were diluted 1:100 in 10 mL of fresh DM2000. The cultures were grown for approximately 3 hours and total RNA was isolated from the mid-log phase cultures with the hot phenol method (Schmitt et al., 1990). Purified total RNA was diluted to 300 ng/μL and treated with TURBO DNase I (Ambion) following the manufacturers' guidelines. The DNase I-treated total RNA was amplified using Phusion® High-fidelity DNA Polymerase (Thermo Fisher) with GFP_F1 and GFP_R1 primers to check for genomic DNA contamination. The DNA-free total RNA was then subjected to first-strand synthesis using SuperScript II (Invitrogen) with the gene-specific primers: GFP_F1 and GFP_R1 following the manufacturers' guidelines. The first-strand cDNA was then amplified using Phusion® High-fidelity DNA Polymerase (Thermo Fisher) with GFP_F1 and GFP_R1. The amplified PCR products were then A-tailed with Taq Polymerase (Bioline), cloned into pGEM-T easy (Promega) and sequenced bi-directionally (Macrogen). The total RNA was also checked for mRNA viability prior to first-strand synthesis with *gstA* specific primers using SuperScript® One-Step RT-PCR System with Platinum® *Taq* DNA Polymerase kit (Invitrogen).

Primers used to detect GFP transcripts were:

GFP_F1:

5′ – AGG AGG ACA GCA ATG CAT CA – 3′

GFP_R1:

5′ – TTA TTT GTA TAG TTC ATC CAT GCC ATG – 3

OD_{595} and in-cell GFP fluorescence measurements

The aforementioned REL606 GFP reporter strains were streaked to single colonies and these single colonies were grown in DM2000 at 37°C for 2 days. The bacterial cultures were diluted 1:100 in 1 mL of fresh DM2000 in a 24 well cell culture plate. OD_{595} and Relative Fluorescent Unit (RFU) measurements were taken every six minutes for 24 hours at 37°C using a FLUOstar Omega Microplate Reader (BMG Labtech) (with shaking between readings at 200 rpm). The excitation wavelength was set to 485 nm, the emission wavelength to 520 nm and the gain to 1000. The gain controls voltage to the detector, making it more or less sensitive based on the signal produced. Pilot tests have shown that when the gain was set to zero, the REL606 GFP reporter systems' RFU readings were beyond the saturation point. All experiments were performed with 6 biological replicates, along with 3 technical replicates per biological replicate.

RNA secondary structure analysis

The RNA sequences of the GFP reporter systems were analysed using Geneious v9.1.3 (Kearse et al., 2012). The RNA secondary structure prediction was performed using the RNA fold feature in Geneious, which utilises the Andronescu energy model and the RNA structure was predicted at 37°C degrees (Andronescu et al., 2007). This energy model utilises a constraint-based parameter estimation algorithm that combines structural and thermodynamic RNA secondary structure data (Andronescu et al., 2007).

SDS-PAGE and western blotting

The GFP reporter system strains were streaked to single colonies and overnight cultures were prepared from the single colonies in DM2000. A 1:100 dilution of the overnight cultures were prepared in 500 mL of fresh DM2000 media. Stationary phase cultures were spun and

the cell pellet was resuspended in Tris Buffer, pH 8.0 and the mixtures were then sonicated. The total soluble protein lysate concentration was determined using a Protein Assay (Bio-Rad). A standard amount of total protein (40 ng) was loaded on a NuPAGE gel (Thermo Fisher), the gel was run at 165 V for 35 mins, and the protein was transferred to a 0.2 µm nitrocellulose membrane (Bio-rad). The membrane was subsequently incubated with Blocking Buffer (1xPBS containing 5% skim milk) for 2 hours and washed with Washing Buffer (1xPBS with 0.05% TWEEN 20). The membrane was immunoblotted with goat Anti-GAPDH (GenScript) in a 1:3000 dilution in Antibody Dilution Buffer (1xPBS supplemented with 1% BSA and 0.05% TWEEN 20) overnight. Subsequently, the membrane was washed and immunoblotted with donkey anti-goat IgG-HRP secondary antibody (Santa Cruz Biotechnology) in a 1:5000 dilution in Antibody Dilution Buffer for 2 hours at room temperature. The membrane was washed and dried before proceeding with the luminol reaction according to the manufacturers' protocol (Santa Cruz Biotechnology). The antibodies were stripped off the membrane with a mild acidic stripping buffer and washed with Washing Buffer prior to immunoblotting with mouse monoclonal Anti-Green Fluorescent Protein (GFP), N-terminal antibody (Sigma) followed by goat anti-mouse IgG-HRP secondary antibody (Santa Cruz Biotechnology) using the aforementioned method.

GFP protein structural modelling

The predicted GFP sequences were used to create a homology model using Phyre2 (Kelley et al., 2015). The GFP homology models were generated using a GFP structure from the PDB (ID:c3ai5A) as the templates for the full-length GFP while the truncated GFP modelled based on two other structures (ID:c4bduC and c3ai5A, respectively). Protein structures were visualised using PyMOL v1.3r1 (Schrödinger, 2015).

Results

RT-PCR confirmed the production of KRS1 and GFP transcripts

To assess whether the KRS1 and GFP genes are producing RNA transcripts, RT-PCR was conducted using primers internal to KRS1 and GFP mRNA. KRS1 and GFP transcripts were detectable in *S. cerevisiae* S288c, including the Slip and Frameshift GFP reporters (Figure 4.3).

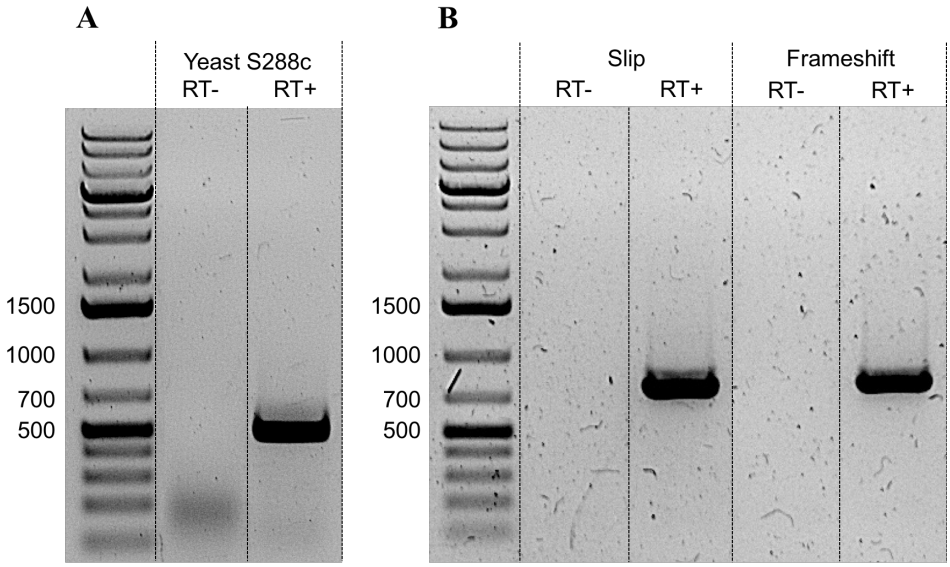

Figure 4.3 | RT-PCR confirms the expression of KRS1 and GFP transcripts. Total RNA was extracted from mid-log phase of *S. cerevisiae* S288c and *E. coli* (Slip and Frameshift) cultures. RT- lanes were processed without a reverse transcriptase reaction in order to detect the presence of genomic DNA. RT+ lanes were processed as per the manufacturer's guidelines which included a reverse transcription step. **A)** RT-PCR with primers corresponding to *S. cerevisiae* S288c KRS1 gene (expected size 499 bp). Lane 1, GeneRuler 1 kb Plus (Thermo Scientific) with size standards indicated; lanes 2 and 3, RT-PCR products for *S. cerevisiae* S288c with RT- controls indicated. **B)** RT-PCR with primers corresponding to GFP (expected product size 759 bp). Lane 1, GeneRuler 1 kb Plus (Thermo Scientific) with size standards indicated; lanes 2-5, RT-PCR products for Slip and Frameshift with RT- controls indicated.

RNA polymerase slippage produces a heterogeneous population of mRNAs

To assess the impact of RNA polymerase slippage at the RNA level, GFP reporter strains that harbour slippage-prone polyA tracts were generated, total RNA was extracted, reverse transcribed, PCR amplified, cloned and sequenced. We observed a population of heterogeneous cDNAs with different lengths of polyA tracts for both the Slip and Frameshift GFP reporter systems (Figure 4.4).

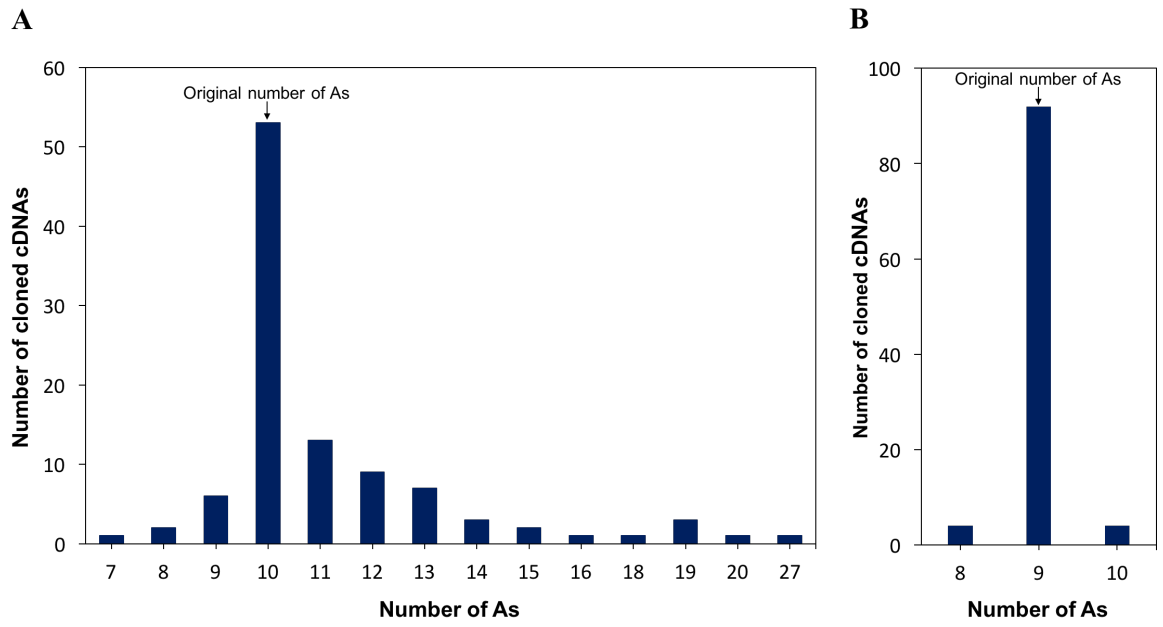

Figure 4.4 | Heterogeneous mRNAs are produced as a result of transcriptional slippage. **A)** The population of cDNAs produced by the Slip GFP construct. A proportion of cDNAs were in-frame, frameshifted with additional or fewer nucleotides compared to the original number of As in the genomic DNA (10 As). **B)** The population of cDNAs produced by the Frameshift GFP construct. While the majority of cDNAs showed no signs of slippage, A small subset of cDNAs were frameshifted with additional or fewer nucleotides compared to the original number of As in the genomic DNA (9 As).

The calculated percentage of slippage on tracts of poly 10A was 48.5% while the percentage of slippage on tracts of poly 9A was about 8% (Table 4.1). The calculated percentage of slippage observed reduced with the reduction in the length of the polyA tract. Further to this, out of the 50 transcripts that showed signs of slippage in the Slip GFP reporter system, 24% of the transcripts were in the correct frame where there was either addition or removal of the number of As in a 10±3n fashion for the 10 As tract. Alternatively, out of the 4 transcripts that showed signs of RNA polymerase slippage in the Frameshift GFP reporter system, 50% of them were in-frame.

Table 4.1 | The number of As observed in slippage-prone poly(A) tracts in *S. cerevisiae* and *E. coli.*

Species	*S. cerevisiae*		*E. coli*			
	KRS1 gDNA	**KRS1 cDNA**	**Slip gDNA**	**Slip cDNA**	**Frameshift gDNA**	**Frameshift cDNA**
7A	0	0	0	1	0	0
8A	0	0	0	2	0	4
9A	0	9	0	6	**50**	**92**
10A	**50**	**82**	**50**	**53**	0	4
11A	0	4	0	13	0	0
12A	0	0	0	9	0	0
13A	0	0	0	7	0	0
14A	0	0	0	3	0	0
15A	0	0	0	2	0	0
16A	0	0	0	1	0	0
17A	0	0	0	0	0	0
18A	0	0	0	1	0	0
19A	0	0	0	3	0	0
20A	0	0	0	1	0	0
27A	0	0	0	1	0	0
Total	50	95	50	103	50	100
Slippage %	0	13.6	0	48.5	0	8
In-frame %	100	86	100	62.1	0	4

Number in bold, the number of tracts with the original number of As in the genomic DNA. Numbers underlined, the number of As in-frame. No slippage was observed at the DNA level for all the samples examined.

We also examined whether the reverse transcriptase used to synthesise the first-strand cDNA was slippage-prone. To test this, we performed RT-PCR using *S. cerevisiae* S288c RNA, and individual cDNAs were subsequently cloned and sequenced. We observed transcript heterogeneity in *S. cerevisiae* S288c KRS1 however, the frequency of the commercial reverse transcriptase (SuperScript II) error was significantly lower than the level of slippage inferred from *E. coli* GFP (Table 4.1) (*P*-value = 0.003, Wilcoxon rank-sum). Direct sequencing of *S. cerevisiae* KRS1 RT-PCR products without cloning showed that either yeast endogenous RNA polymerase, SuperScript II or DNA polymerases (*S. cerevisiae* endogenous polymerase and Phusion) were slipping on the polyA tracts in *S. cerevisiae* KRS1 gene (Supplementary Figure 4.1).

DNA polymerases tested are not slippage-prone

To test whether the commercial DNA polymerase used (Phusion) is slippage-prone, we amplified yeast KRS1 gene and sequenced the PCR product. Additionally, to test for slippage of the DNA polymerase used for Sanger sequencing, we extracted and sequenced the Slip and Frameshift GFP plasmids. These control experiments showed that the DNA polymerases used in this experiment were not slippage-prone (Figure 4.5 and Table 4.1) where the each peak showed high confidence in base calling.

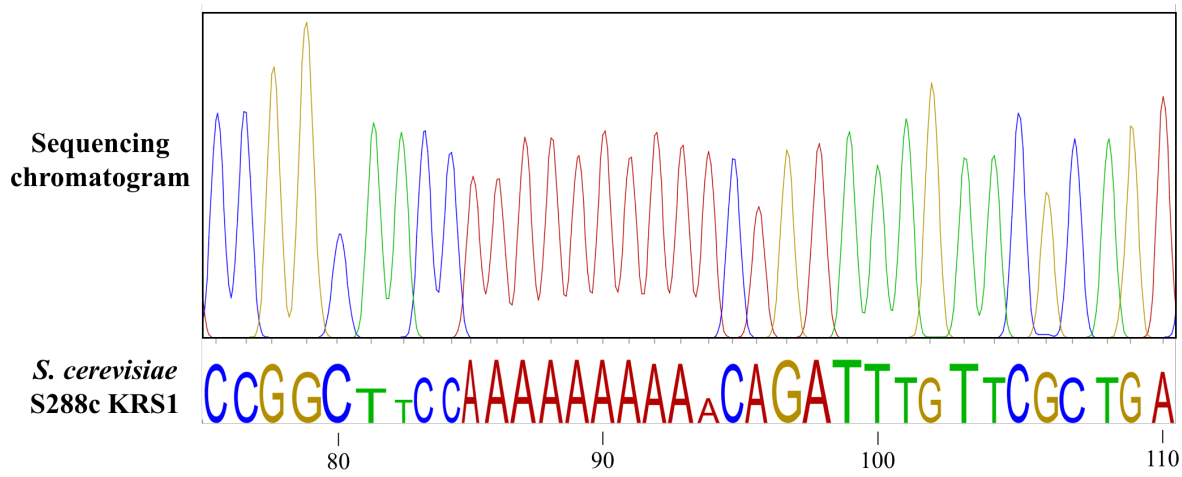

Figure 4.5 | Chromatogram showing no slippage at the DNA level. *S. cerevisiae* KRS1 gene was amplified with Phusion® High-fidelity DNA Polymerase (Thermo Fisher) and showed no signs of slippage where each peak showed high confidence in base calling. The sequence logo below the chromatogram shows the level of confidence each base was called, where the higher the alphabet, the higher confidence, and vice versa.

RNA polymerase slippage affects cell growth

To assess the effect of polyA tracts on gene expression, we measured the OD_{595} and Relative Fluorescence Units (RFU) of GFP every 6 minutes for 24 hours on a plate reader. The growth trend differs among the strains, where we observed a longer lag phase for No Slip, and the stationary phase of Frameshift peaks at approximately OD_{595} of 0.38, whereas the other 3 strains (Vector, Slip and No Slip) peaked above an OD_{595} of 0.42 (Figure 4.6). In order to determine whether RNA polymerase slippage affects growth, we calculated the minimum doubling time taken over a 30-minute interval. The calculated doubling time was 59.0±1.6 min and 59.9±0.5 min for the Vector and No Slip GFP reporter strains respectively (Figure 4.6). The calculated doubling times for the Slip and Frameshift GFP reporter systems were lower compared to Vector and No Slip at 47.2±0.9 min and 52.3±1.2 min respectively (*P-value* = 5.14×10^{-10}, Kruskal-Wallis).

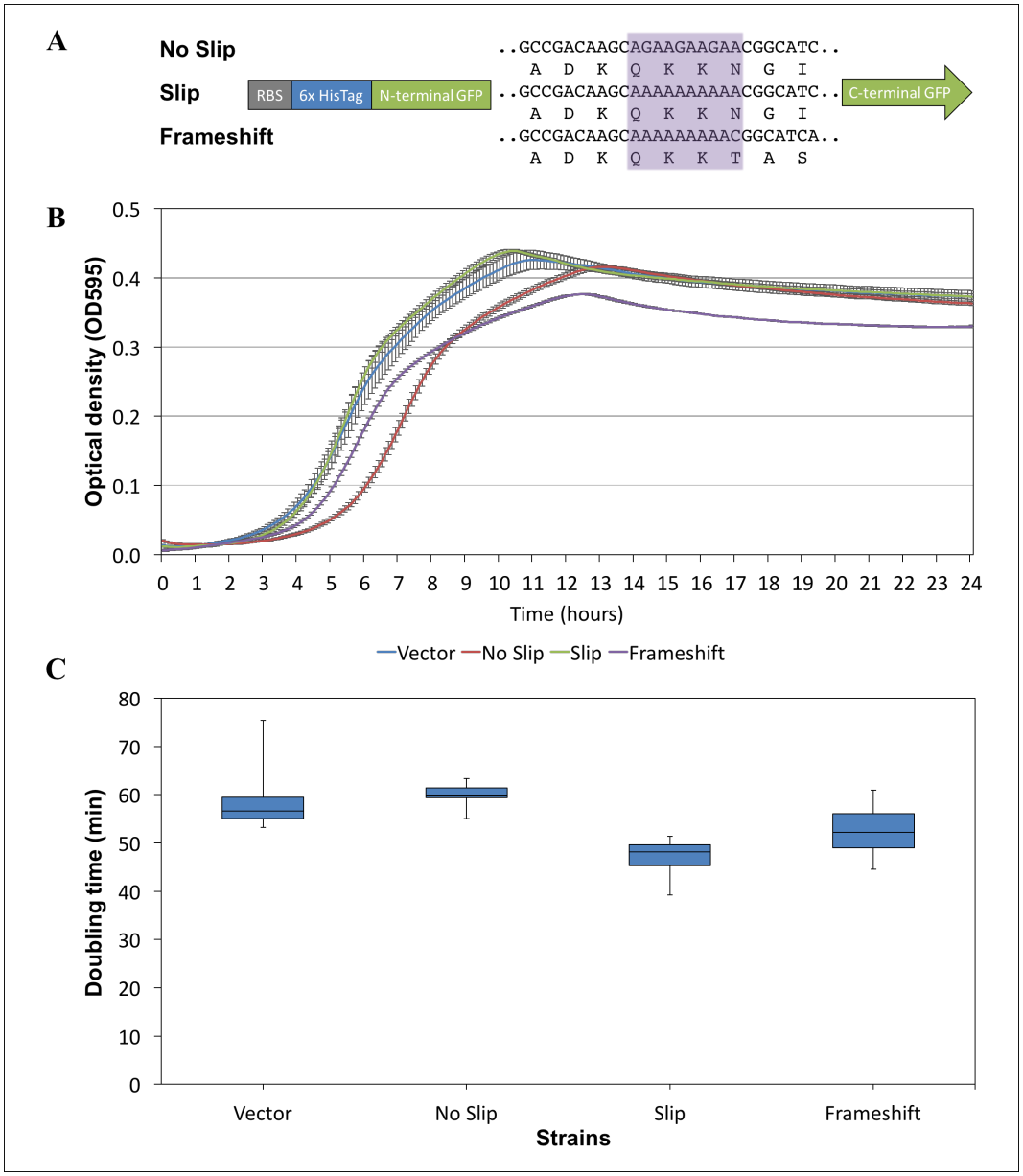

Figure 4.6 | Homopolymeric tracts of A affects growth rates. REL606 strains carrying the GFP reporter systems including a control (REL606 + pGEM-T easy, termed Vector) were grown in DM2000 at 37°C with shaking, and the optical density (OD595) was collected every 6 minutes for 24 hours. **A)** The GFP reporter strains, consisting of plasmid-encoded GFP genes with slippage-prone polyA tracts as indicated by the purple box. **B)** The OD595 was monitored for 24 hours. Points are shown as $\bar{X}$±SD, n=18. **C)** The calculated doubling time of Vector, No Slip, Slip and Frameshift were 59.0±1.6 min, 59.9±0.5, 47.2±0.9 min and 52.3±1.2 min respectively (*P-value* = 5.14 x 10^{-10}, Kruskal-Wallis). Doubling times are shown as $\bar{X}$±SE, n=18.

RNA polymerase slippage reduced GFP production

To determine the effect of RNA polymerase slippage on protein production, we measured GFP production as Relative Fluorescence Unit (RFU) of the aforementioned GFP reporter systems. We observed the production of fluorescence for Vector (REL606 carrying empty pGEM-T easy plasmids) in the absence of a GFP gene (Figure 4.7). The level of fluorescence emitted from Vector was 4457±295 RFU while the RFU levels for No Slip, Slip and Frameshift GFP reporter systems were 10865±142, 6348±97, and 5513±48 respectively ($\bar{X}$±SE, n=18).

To determine the net GFP production, the RFUs of the GFP reporter systems were corrected against Vector RFU by subtracting the RFU values of the GFP reporter systems with the RFU level of Vector (Figure 4.7). The Vector corrected RFU of No Slip was 6408±142 while Slip and Frameshift recorded RFUs of 1892±97 and 1057±48 respectively ($\bar{X}$±SE, n=18). The production of GFP fluorescence reduced with the introduction of slippage-prone polyA tracts (Figure 4.7). We observed approximately 3-fold and 6-fold less GFP production for Slip and Frameshift compared to the No Slip control (*P*-value = 5.0×10^{-4}, Kruskal-Wallis).

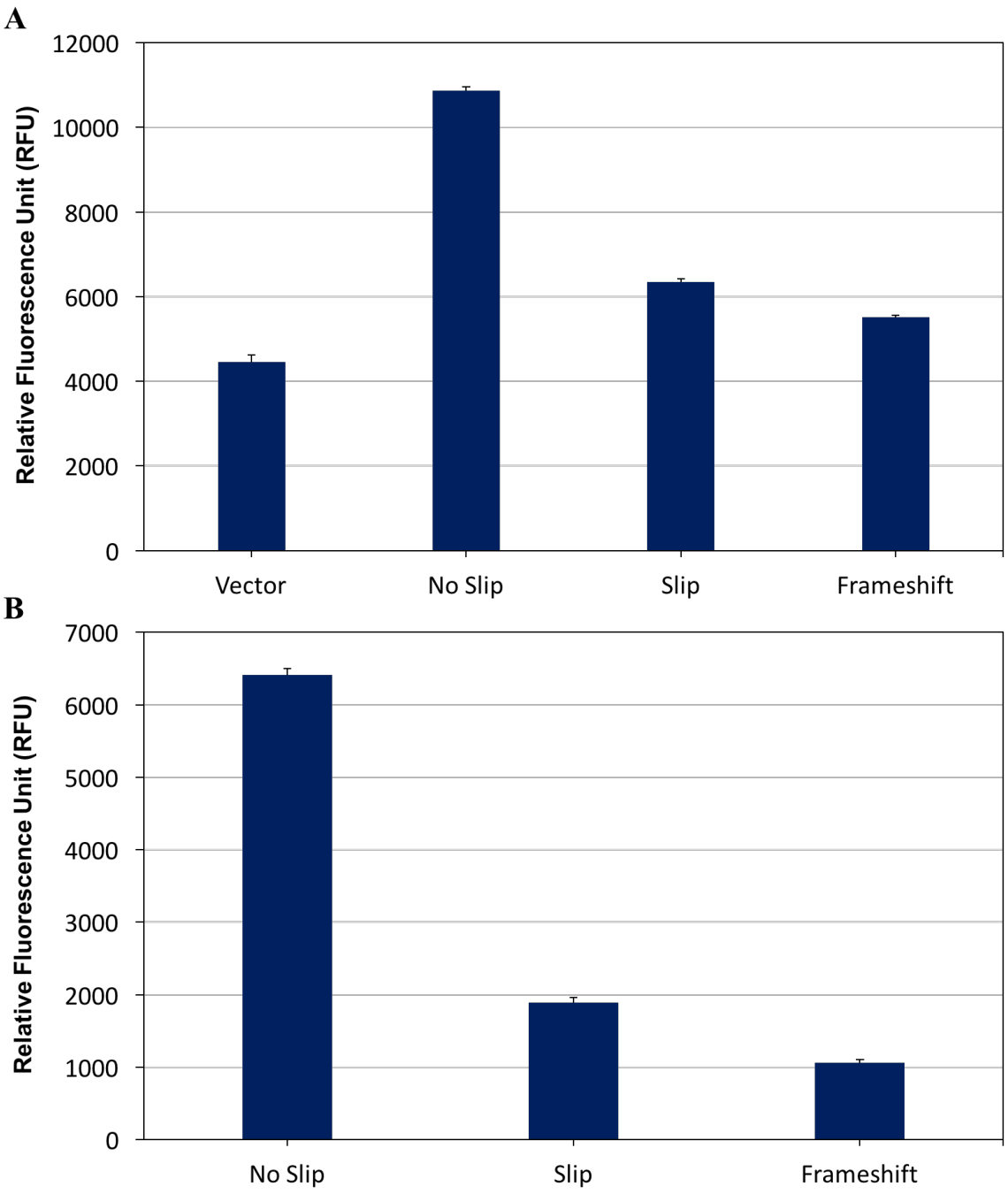

Figure 4.7 | Measurements of end-point GFP production in minimal media. The RFU was measured after 24 hours of incubation in Davis Minimal supplemented with 0.2% glucose. The GFP fluorescence (ex. 485 nm, em. 520 nm) was measured as Relative Fluorescence Units (RFU), and the gain was set at 1000. **A)** The RFU of the Vector, No Slip, Slip, and Frameshift reporter systems were 4457±295, 10865±142, 6348±97, and 5513±48 respectively ($\bar{X}$±SE, n=18). **B)** The Vector corrected RFUs for No Slip, Slip, Frameshift and Double Slip were 6408±142, 1892±97, and 1057±48 respectively ($\bar{X}$±SE, n=18).

Slippage-prone polyA tracts may be prone to ribosomal frameshifting

Homopolymeric A/T tracts are also prone to ribosomal frameshifting (Anikin et al., 2010; Sharma et al., 2014), and to determine whether ribosomal frameshifting may have contributed to the observed reduction of GFP production, we analysed the RNA secondary structures of the GFP reporter systems. When compared to the *dnaX* gene secondary structure, our GFP reporter gene does not contain a Shine-Dalgarno (SD)-like sequence prior to the slippage-prone A.AAA.AAC motif, a spacer and a 3' stimulatory RNA fold structure (Figure 4.8).

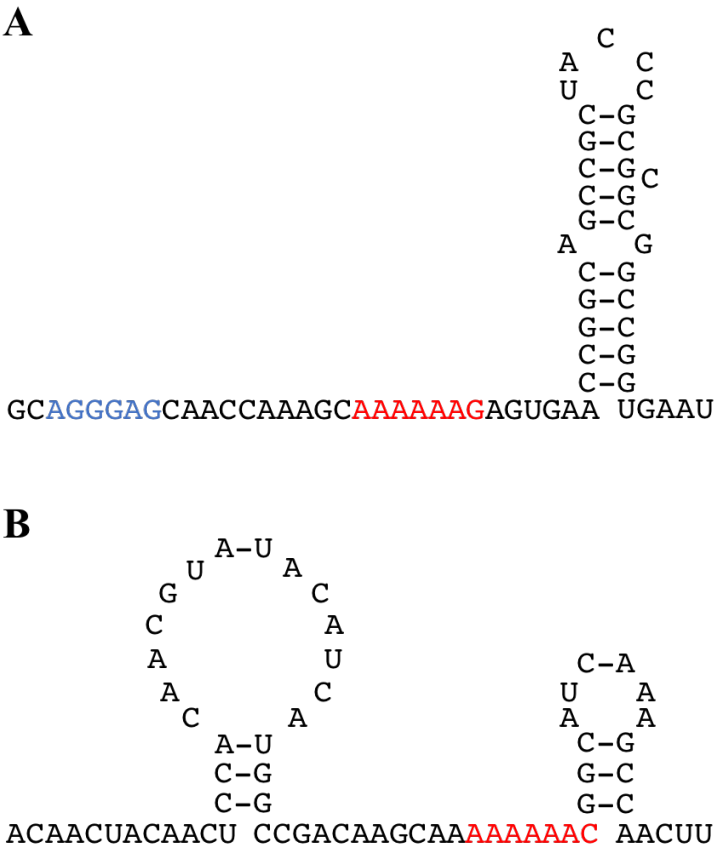

Figure 4.8 | Motifs required for efficient ribosomal frameshifting. The blue letters represent SD-like sequence while the red letters represent a slippery motif. **A)** The *dnaX* gene in *E. coli* which has been shown to be regulated by ribosomal frameshifting consists of a SD-like region prior to the slippage-prone motif, a spacer and a 3' stimulatory pseudoknot. **B)** Our GFP reporter consists only of a slippage-prone motif followed by a stem-loop.

Protein expression was not completely eliminated by slippage

To determine whether full-length GFP proteins are produced from the population of GFP mRNAs with varying length and reading frames (Table 4.1), we extracted total protein from the aforementioned *E. coli* strains and conducted western blot using monoclonal N-terminal anti-GFP antibody. As expected, we did not observe any bands that corresponded to full-length GFP for the Vector control. We observed a single bright band that corresponded to the predicted full-length GFP, ~28 kDa for No Slip (Figure 4.9). Furthermore, we observed a bright band of ~28 kDa and a faint band of ~22 kDa that corresponded to full-length and predicted truncated N-terminal GFP for Slip (Supplementary Table 4.1) (Figure 4.9). Strikingly, 3 bright bands of ~28 kDa, ~ 27 kDa and ~22 kDa bands were observed for Frameshift.

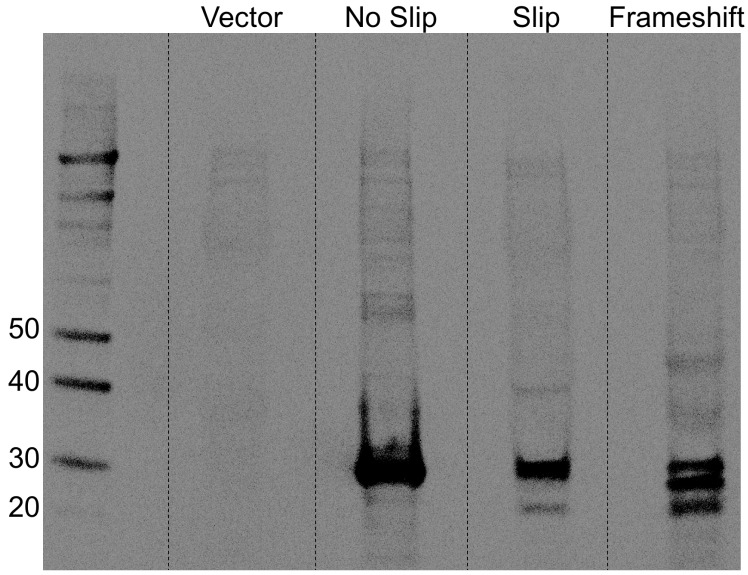

Figure 4.9 | Western blot of total soluble protein probed with anti-GFP antibody. Total soluble protein extracted from stationary phase cultures of *E. coli* REL606 with the aforementioned GFP reporter systems were immunoblotted with monoclonal N-terminal anti-GFP antibody. Lane 1, Novex® Sharp Pre-stained Protein Standard (Thermo Scientific) with size standards indicated. Lane 2-5, total soluble protein from the Vector, No Slip, Slip and Frameshift GFP reporter systems. Full-length GFP was observed for all the GFP strains.

Production of GFP protein with different lengths and structures

We previously observed the production of three different GFP protein products using western blot (Figure 4.9). To test whether the western blot results are consistent with RNA polymerase slippage, we performed structural modelling of the predicted protein products (Supplementary Table 4.1). We observed three major GFP structures: full-length GFP (when the number of As is equal to the templated 10 or 10±3n), and two variants of truncated GFP (when the number of As does not equal to the templated 10 or 10±3n) (Figure 4.10).

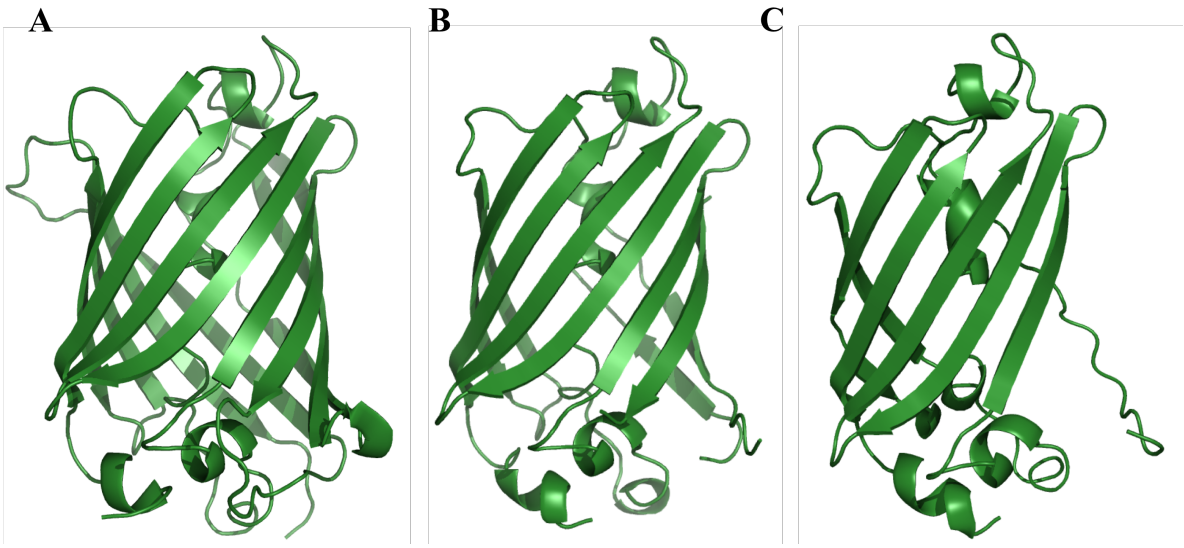

Figure 4.10 | The predicted structures of GFP proteins. Phyre2 was used to generate structural homology models for the predicted GFP upon RNA polymerase slippage. The structures were modelled based on the template of which the GFP sequence has the highest identity to. **A)** The predicted full-length GFP model based on c3ai5A consisted of 11 antiparallel beta strands and an alpha helix that runs through the centre of the GFP TIM barrel(Ormö et al., 1996; Yang et al., 1996). **B)** A truncated GFP model based on c4bduC which consisted of 9 beta strands and an alpha helix. **C)** The other version of a truncated GFP model based on c3ai5A, consisting of 6 beta strands and an alpha helix. Images generated in PyMOL (Schrödinger, 2015).

Discussion

Homopolymeric tracts are unstable and are prone to indels as a result of misalignment of DNA strands, resulting in expansion or contraction of the DNA strand (Canceill et al., 1999; Streisinger et al., 1966). Genes that contain these tracts are susceptible to slippage-type editing, an enzymatic error that results in a heterogeneous pool of messenger RNA transcripts with varying reading frames and length (Tamas et al., 2008). Translation of these transcripts would generally produce aberrant, truncated or non-functional proteins that could be detrimental to the cell (Baranov et al., 2005). The deleterious consequent of homopolymeric tracts may explain the under-representation of these tracts in most bacteria genes and their non-random location within genes where they persist (Dechering et al., 1998; van Passel and Ochman, 2007). Although these homopolymeric tracts occur infrequently in genomic DNA, they has been suggested to play a regulatory role in gene expression (Han and Turnbough, 1998; Liu et al., 1994; Qi and Turnbough Jr, 1995; Tu and Turnbough, 1997).

In Chapter 3, we showed that these inherently disadvantageous slippage-prone tracts may readily emerge under conditions favouring drift. We also demonstrated that in some cases RNA polymerase slippage may result in a loss of fitness. Here, we gauged on the impact of slippage-type editing on gene expression and protein production. To investigate the impact of slippage on gene expression, we introduced polyA tracts into green fluorescence protein (GFP) reporter systems, isolated the RNA, reverse transcribed the RNA, and sequenced the cloned cDNAs. We observed heterogeneity of the GFP RNA transcripts for both the Slip and Frameshift GFP reporter systems (Figure 4.4). The observed transcript heterogeneity is consistent with previous results presented on the effect of slippage on RNA transcript heterogeneity (Tamas et al., 2008; Wagner et al., 1990).

We observed a higher frequency of slippage on repeats of 10 As (Slip) compared to 9 As (Frameshift), where approximately 49% of the cloned cDNA sequences showed length heterogeneity that differs from the original DNA template, whereas only 8% of RNA polymerase slippage was recorded for 9 As (Table 4.1). This result may be explained by the length of the polyA tract, as RNA polymerase slippage relies on the length and stability of the DNA-RNA duplex within the polymerase (Parks et al., 2014; Zhou et al., 2013). Studies on the DNA-RNA hybrid in maintaining the correct registry of the ternary elongation complex (TEC) showed that the DNA-RNA hybrid is 8-9 bp in length in *E. coli* (Nudler et al., 1997). Although the suggested minimum length of polyA/T that induces RNA polymerase was 9A/T (Gordon et al., 2013; Larsen et al., 2000), we observed low levels of slippage in our GFP reporter system bearing tracts of 9 As (Table 4.1) in our experiment. Slippage has been shown to occur more frequently on longer tracts of polyA/T and slippage efficiency reduces with shorter runs (Penno et al., 2015; Uptain et al., 1997; Wagner et al., 1990). The shorter length of the polyA tracts could account for the lower slippage frequency observed for the tracts of 9 As compared to 10 As in our experiment.

We also noted that within the subset of transcripts that showed slippage, 24% and 50% of the transcripts were in-frame (when the number of As is equal to the templated 10 or 10±3n) for the Slip and Frameshift GFP reporter systems respectively. These results suggest that slippage-type editing may help rescue gene expression of genes bearing frameshifted polyA tracts but reduce gene expression efficiency of genes with in-frame polyA tracts (discussed below).

As we previously observed GFP RNA transcript heterogeneity (Figure 4.4), we then asked whether the observed results were artefacts of DNA polymerase, and/or reverse transcriptase

slippage. DNA polymerase slippage has been observed in *S. cerevisiae* (Gragg et al., 2002), *E. coli* (Canceill et al., 1999) and commercially available DNA polymerases (Clarke et al., 2001). We performed multiple control experiments to determine whether *E. coli* endogenous DNA polymerase and commercial enzymes used were slippage-prone. Control experiments indicated that the observed length heterogeneity did not reflect DNA replication, PCR, cloning, or DNA sequencing artefacts (Table 4.1). Moreover, we conducted the same experiment using *S. cerevisiae* S288c containing a poly 10A tract and our results demonstrated that the commercial SuperScript®II enzyme used was slippage-prone, however, the frequency of error was significantly lower than slippage inferred in *E. coli* GFP gene with 10A tracts (Table 4.1) (*P*-value = 0.003, Wilcoxon rank-sum). No heterogeneity was observed at the DNA level, suggesting that endogenous DNAP and DNA polymerases used in cDNA amplification and sequencing did not contribute to the changes observed in the transcripts. We also addressed the potential for compensatory mutations in the genomic DNA sequence (Szamecz et al., 2014), but DNA sequencing of the plasmids bearing polyA tracts (pGEM::Slip and pGEM::Frameshift) revealed no additional indels to the gene (Table 4.1). We, therefore, conclude that the observed cDNA heterogeneity reflects the underlying variation in transcript sequences and is attributable to slippage by endogenous *E. coli* RNA polymerase.

Transcript heterogeneity is predicted to impact translation as some proportion of mRNAs are frameshifted, some exhibit codon addition or deletion while a subset carries the correct sequence as shown in our results. We have previously shown that slippage-type editing resulted in transcript heterogeneity, and to assess the effect of slippage on protein production, we measured the relative fluorescence units (RFU) produced by GFP reporter strains. We observed a reduction of RFU levels in Slip and Frameshift constructs compared to the No

Slip control GFP reporter system (Figure 4.7). In the absence of RNA polymerase slippage, we would expect similar levels of RFU produced by both the No-Slip and Slip reporter system as they are both in the same frame. The synonymous changes to the sequence from AG.AAG.AAA.AA to AA.AAA.AAA.AA should not affect the polypeptide sequence as both sequences are subsequently translated to QKKN, yet we observed approximately 3-fold less RFU produced by the Slip reporter. Additionally, we observed approximately 6-fold less RFU produced by the Frameshift reporter compared to No Slip, and this result is consistent with the aforementioned observed RNA transcripts, where only 8% of the transcript showed signs of slippage (Table 4.1). These results suggest that RNA polymerase slippage results in the production of a heterogeneous population of mRNAs and that subsequent translation of these mRNAs results in a proportion of full-length and truncated GFP proteins, consistent with the observed reduction of RFU level.

We have demonstrated that RNA polymerase slippage reduced GFP production (as measured in RFUs), however, ribosomal frameshifting has also been documented to occur on polyA/T tracts in *E. coli.* (Gurvich et al., 2003; Sharma et al., 2014). To examine whether ribosomal frameshifting may have contributed to the reduction of GFP production in our experiment, we analysed the RNA secondary structure of our GFP reporter systems and compared the structure to *E. coli dnaX* that is prone to ribosomal frameshifting. We did not observe several fundamental motifs required for efficient ribosomal frameshifting in our reporter system, including a Shine-Dalgarno (SD)-like sequence prior to the slippage-prone motif followed by a spacer and a 3' stimulatory pseudoknot (Atkins et al., 2016; Namy et al., 2006), when we compared the secondary RNA structure of our GFP reporter to *E. coli dnaX* (Figure 4.8). Further to this, a study combining bioinformatics and experimental approaches on identifying efficient ribosomal frameshifting motifs in *E. coli* showed that the A.AAA.AAC motif in our

GFP reporter was inefficient for ribosomal frameshifting to occur in *E. coli* (Sharma et al., 2014). These observations suggest that the reduction of GFP production observed in our experiment was a result of RNA polymerase slippage and not ribosomal frameshifting. Although the RNA secondary structure comparison suggested that ribosomal frameshifting was not responsible for the low levels of GFP observed, we have not provided any experimental evidence to support this prediction. To test this, we could perform ribosome profiling (Michel and Baranov, 2013) to monitor and identify whether ribosome frameshifting occurs in our GFP reporter systems.

Having shown that a proportion of the heterogeneous mRNAs produced as a result of RNA polymerase slippage could serve as substrates for translation of full-length proteins, we assessed whether full-length functional GFPs are expressed. Western blots on total protein extracts using N-terminal monoclonal Anti-Green Fluorescent Protein (GFP) antibody gave a band of ~28 kDa (Figure 4.9), corresponding to the predicted molecular weight of the full-length GFP for all the reporter systems (Supplementary Table 4.1). Interestingly, a band of ~22 kDa was observed for both the Slip and Frameshift reporter systems while an additional band of ~27 kDa was observed for Frameshift (Figure 4.9). These bands corresponded to the predicted sizes of truncated GFP upon the insertion or removal of As in the mRNAs (Supplementary Table 4.1). Further to this, we also observed less full-length GFP (based on brightness of the western blot bands) being produced in the Slip and Frameshift GFP reporter systems, given that an equal amount of total protein was loaded. This indicates that RNA polymerase slippage produced a proportion of transcripts that were in the correct frame (when the number of A=10 or 10±3n), and a proportion of frameshifted transcripts. However, western blot alone is not sufficient to demonstrate that these bands were indeed the predicted GFP proteins and this method only provides a semi-quantitative measurement of GFP

produced. In order to investigate whether these bands are indeed the protein products that we predicted, we could perform mass spectrometry. The shotgun (or discovery) proteomic analysis would enable us to identify whether the bands observed in our western blot correspond to the predicted GFP peptide sequences (Lu et al., 2009). Alternatively, the selected reaction monitoring (SRM) analysis would enable us to quantify the amount of GFP produced for each GFP reporter strain (Domon and Aebersold, 2006, 2010).

We next examined whether the predicted GFP sequences could form functional GFP products by modelling the predicted GFP sequences. The full-length GFP including elongated and truncated GFP (where the number of As is 10±3) consists of the fundamental 11 antiparallel beta strands and an alpha-helix that runs through the centre, in the middle of the TIM barrel that is the chromophore (Ormö et al., 1996; Yang et al., 1996) (Figure 4.10). We also modelled the predicted truncated GFP and we observed two major variations of truncated GFP. The first structure consisted of 9 beta strands and an alpha helix while the second structure consisted of 6 beta strands and an alpha helix (Figure 4.10). Deletion of more than the N-terminal methionine residue or more than seven amino acids from the C-terminal has been shown to lead to a total loss of fluorescence in addition to the loss of the characteristic absorption spectrum of the intact fluorophore, as the GFP beta-can structure cannot be formed (Yang et al., 1996). We observed bands that corresponded to the predicted truncated GFP in our western blot (Figure 4.9) however, we would not expect to see fluorescence from this truncated GFP. This might account for the reduced fluorescence observed in the Slip and Frameshift reporters, as we are seeing a proportion of truncated protein, compared to the sole production of full-length GFP in the No Slip construct, and the oberservation of multiple bands is consistent with what we would expect if slippage-type editing were occurring

(Figure 4.7). We can, therefore, conclude that slippage-type editing occurred, resulting in the production of a proportion of full-length and truncated GFP transcripts.

We have shown that slippage-type editing can provide the advantage of rescuing the functionality of genes with frameshifted polyA tracts (Figure 4.7 and 4.10). Further to this, slippage-type editing has also been shown to produce multiple functional proteins from one gene (Baranov et al., 2005; Larsen et al., 2000; Penno et al., 2006; Schurig et al., 1995). The utilisation of slippage-type editing in the synthesis of additional proteins from a single gene has also been well-documented in sendai virus (Hausmann et al., 1999) and Ebola virus (EBOV) (Lee and Saphire, 2009; Sanchez et al., 1996; Volchkov et al., 1995; Volchkov et al., 2001) (see Chapter 1). The generation of multiple products from a single gene was thought to provide variability and optimise the use of the limited genetic information in response to changes in environmental and growth conditions (Anikin et al., 2010). However, we did not observe the production of elongated or additional GFP proteins as no novel gene functions are anticipated for the alternate forms of GFP in our study.

To summarise, homopolymeric tracts, particularly stretches of As or Ts, are prone to enzymatic slippage producing RNA outputs that do not reflect the DNA template. Slippage-type editing can restore the open reading frame by correcting the reading frame at the mRNA level in cases where the open reading frame has acquired a frameshift mutation at the gene level. This, in turn, leads to the expression of full-length functional proteins, however, RNA polymerase slippage may also reduce expression efficiency of in-frame genes.

Supplementary

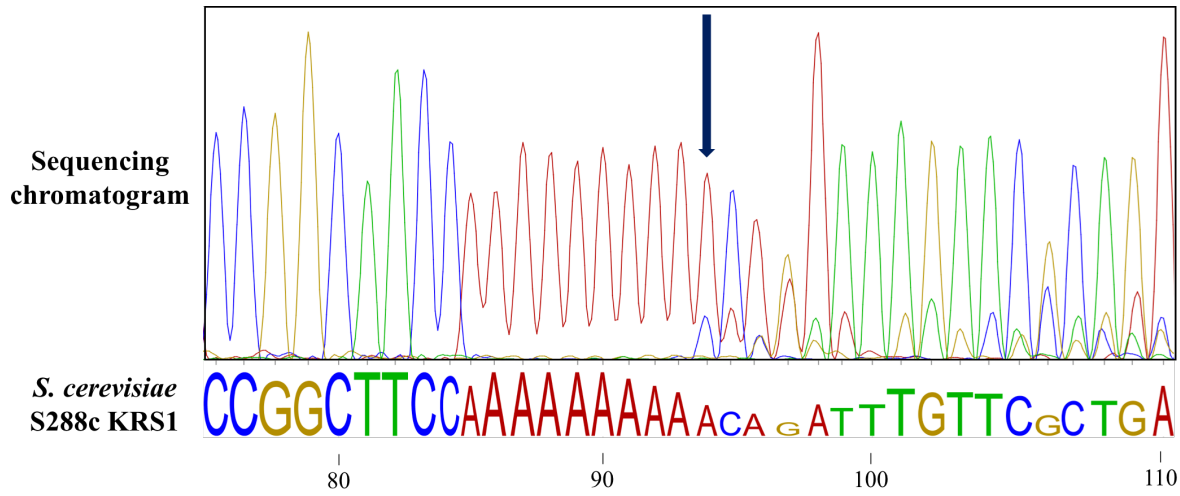

Figure 4.1 | Chromatogram showing signs of slippage. The blue arrow points to the region where slippage occurs generating mRNAs of various lengths as shown by the multiple peaks observed in one base position at the 10th A (position 94) and every base downstream from the blue arrow. The sequence logo below the chromatogram shows the level of confidence each base was called where the higher the alphabet, the higher confidence, and vice versa.

Table 4.1 | The predicted molecular weight of GFP proteins translated from mRNA transcripts with varying lengths of A

Number of As	Molecular weight (kDa)
1	27.82
2	21.62
3	26.26
4	27.93
5	21.74
6	26.39
7	28.05
8	21.87
9	26.52
10	28.18
11	22.00
12	26.65
13	28.31
14	22.13
15	26.77
16	28.44
17	22.26
18	26.90
19	28.56
20	22.39

The molecular weights were predicted using ExPASy (Gasteiger et al., 2005)

References

Andronescu, M., Condon, A., Hoos, H.H., Mathews, D.H., and Murphy, K.P. (2007). Efficient parameter estimation for RNA secondary structure prediction. Bioinforma. Oxf. Engl. *23*, i19-28.

Anikin, M., Molodtsov, V., Temiakov, D., and McAllister, W.T. (2010). Transcript slippage and recoding. In Recoding: Expansion of Decoding Rules Enriches Gene Expression, J.F. Atkins, and R.F. Gesteland, eds. (Springer New York), pp. 409–432.

Atkins, J.F., Loughran, G., Bhatt, P.R., Firth, A.E., and Baranov, P.V. (2016). Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. gkw530.

Baranov, P.V., Hammer, A.W., Zhou, J., Gesteland, R.F., and Atkins, J.F. (2005). Transcriptional slippage in bacteria: distribution in sequenced genomes and utilization in IS element gene expression. Genome Biol. *6*, R25.

Canceill, D., Viguera, E., and Ehrlich, S.D. (1999). Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. J. Biol. Chem. *274*, 27481–27490.

Clarke, L.A., Rebelo, C.S., Gonçalves, J., Boavida, M.G., and Jordan, P. (2001). PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Mol. Pathol. *54*, 351–353.

Dechering, K.J., Konings, R.N.H., Cuelenaere, K., and Leunissen, J.A.M. (1998). Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. Nucleic Acids Res. *26*, 4056–4062.

Dinman, J.D. (2012a). Mechanisms and implications of programmed translational frameshifting. Wiley Interdiscip. Rev. RNA *3*, 661–673.

Dinman, J.D. (2012b). Control of gene expression by translational recoding. Adv. Protein Chem. Struct. Biol. *86*, 129–149.

Domon, B., and Aebersold, R. (2006). Mass Spectrometry and Protein Analysis. Science *312*, 212–217.

Domon, B., and Aebersold, R. (2010). Options and considerations when selecting a quantitative proteomics strategy. Nat. Biotechnol. *28*, 710–721.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R., and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In The Proteomics Protocols Handbook, J. Walker, ed. (Humana Press), pp. 571–607.

Gordon, A.J.E., Satory, D., Halliday, J.A., and Herman, C. (2013). Heritable change caused by transient transcription errors. PLOS Genet *9*, e1003595.

Gragg, H., Harfe, B.D., and Jinks-Robertson, S. (2002). Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in *Saccharomyces cerevisiae.* Mol. Cell. Biol. *22*, 8756–8762.

Gurvich, O.L., Baranov, P.V., Zhou, J., Hammer, A.W., Gesteland, R.F., and Atkins, J.F. (2003). Sequences that direct significant levels of frameshifting are frequent in coding regions of *Escherichia coli*. EMBO J. *22*, 5941–5950.

Han, X., and Turnbough, C.L. (1998). Regulation of *carAB* expression in *Escherichia coli* occurs in part through UTP-sensitive reiterative transcription. J. Bacteriol. *180*, 705–713.

Han, X., and Turnbough, C.L. (2014). Transcription start site sequence and spacing between the −10 Region and the start site affect reiterative transcription-mediated regulation of gene expression in *Escherichia coli*. J. Bacteriol. *196*, 2912–2920.

Hausmann, S., Garcin, D., Delenda, C., and Kolakofsky, D. (1999). The versatility of paramyxovirus RNA polymerase stuttering. J. Virol. *73*, 5568–5576.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinforma. Oxf. Engl. *28*, 1647–1649.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858.

Koch, A.L. (2004). Catastrophe and what to do about it if you are a bacterium: The importance of frameshift mutants. Crit. Rev. Microbiol. *30*, 1–6.

Larsen, B., Wills, N.M., Gesteland, R.F., and Atkins, J.F. (1994). rRNA-mRNA base pairing stimulates a programmed -1 ribosomal frameshift. J. Bacteriol. *176*, 6842–6851.

Larsen, B., Wills, N.M., Nelson, C., Atkins, J.F., and Gesteland, R.F. (2000). Nonlinearity in genetic decoding: Homologous DNA replicase genes use alternatives of transcriptional slippage or translational frameshifting. Proc. Natl. Acad. Sci. *97*, 1683–1688.

Lee, J.E., and Saphire, E.O. (2009). Ebolavirus glycoprotein structure and mechanism of entry. Future Virol. *4*, 621–635.

Linton, M.F., Raabe, M., Pierotti, V., and Young, S.G. (1997). Reading-frame restoration by transcriptional slippage at long stretches of adenine residues in mammalian cells. J. Biol. Chem. *272*, 14127–14132.

Liu, C., Heath, L.S., and Turnbough, C.L. (1994). Regulation of *pyrBI* operon expression in *Escherichia coli* by UTP-sensitive reiterative RNA synthesis during transcriptional initiation. Genes Dev. *8*, 2904–2912.

Lu, B., Xu, T., Park, S.K., and Yates, J.R. (2009). Shotgun protein identification and quantification by mass spectrometry. Methods Mol. Biol. Clifton NJ *564*, 261–288.

Macdonald, L.E., Zhou, Y., and McAllister, W.T. (1993). Termination and slippage by bacteriophage T7 RNA polymerase. J. Mol. Biol. *232*, 1030–1047.

Meyerovich, M., Mamou, G., and Ben-Yehuda, S. (2010). Visualizing high error levels during gene expression in living bacterial cells. Proc. Natl. Acad. Sci. *107*, 11543–11548.

Michel, A.M., and Baranov, P.V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley Interdiscip. Rev. RNA *4*, 473–490.

Namy, O., Moran, S.J., Stuart, D.I., Gilbert, R.J.C., and Brierley, I. (2006). A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. Nature *441*, 244–247.

Ninio, J. (1991). Connections between translation, transcription and replication error-rates. Biochimie *73*, 1517–1523.

Nudler, E., Mustaev, A., Goldfarb, A., and Lukhtanov, E. (1997). The RNA–DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. Cell *89*, 33–41.

Ormö, M., Cubitt, A.B., Kallio, K., Gross, L.A., Tsien, R.Y., and Remington, S.J. (1996). Crystal structure of the *Aequorea victoria* green fluorescent protein. Science *273*, 1392–1395.

Orsi, R.H., Bowen, B.M., and Wiedmann, M. (2010). Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. BMC Genomics *11*, 102.

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. Nat. Genet. *31*, 69–73.

Pál, C., and Hurst, L.D. (2000). The evolution of gene number: are heritable and non-heritable errors equally important? Heredity *84*, 393–400.

Paramban, R.I., Bugos, R.C., and Su, W.W. (2004). Engineering green fluorescent protein as a dual functional tag. Biotechnol. Bioeng. *86*, 687–697.

Parks, A.R., Court, C., Lubkowska, L., Jin, D.J., Kashlev, M., and Court, D.L. (2014). Bacteriophage λ N protein inhibits transcription slippage by *Escherichia coli* RNA polymerase. Nucleic Acids Res. *42*, 5823–5829.

van Passel, M.W.J., and Ochman, H. (2007). Selection on the genic location of disruptive elements. Trends Genet. TIG *23*, 601–604.

Penno, C., Sansonetti, P., and Parsot, C. (2005). Frameshifting by transcriptional slippage is involved in production of MxiE, the transcription activator regulated by the activity of the type III secretion apparatus in *Shigella flexneri*. Mol. Microbiol. *56*, 204–214.

Penno, C., Hachani, A., Biskri, L., Sansonetti, P., Allaoui, A., and Parsot, C. (2006). Transcriptional slippage controls production of type III secretion apparatus components in *Shigella flexneri*. Mol. Microbiol. *62*, 1460–1468.

Penno, C., Sharma, V., Coakley, A., O'Connell Motherway, M., van Sinderen, D., Lubkowska, L., Kireeva, M.L., Kashlev, M., Baranov, P.V., and Atkins, J.F. (2015). Productive mRNA stem loop-mediated transcriptional slippage: Crucial features in common with intrinsic terminators. Proc. Natl. Acad. Sci. *112*, E1984–E1993.

Qi, F., and Turnbough Jr, C.L. (1995). Regulation of *codBA* operon expression in *Escherichia coli* by UTP-dependent reiterative transcription and UTP-sensitive transcriptional start site switching. J. Mol. Biol. *254*, 552–565.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226.

Ratinier, M., Boulant, S., Combet, C., Targett-Adams, P., McLauchlan, J., and Lavergne, J.-P. (2008). Transcriptional slippage prompts recoding in alternate reading frames in the hepatitis C virus (HCV) core sequence from strain HCV-1. J. Gen. Virol. *89*, 1569–1578.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, N. Y.: Cold Spring Harbor Laboratory Pr).

Sanchez, A., Trappier, S.G., Mahy, B.W., Peters, C.J., and Nichol, S.T. (1996). The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proc. Natl. Acad. Sci. *93*, 3602–3607.

Schaaper, R.M. (1993). Base selection, proofreading, and mismatch repair during DNA replication in *Escherichia coli.* J. Biol. Chem. *268*, 23762–23765.

Schmitt, M.E., Brown, T.A., and Trumpower, B.L. (1990). A rapid and simple method for preparation of RNA from *Saccharomyces cerevisiae*. Nucleic Acids Res. *18*, 3091–3092.

Schrödinger, L. (2015). The PyMOL molecular graphics systems (Schrödinger, LLC).

Schurig, H., Beaucamp, N., Ostendorp, R., Jaenicke, R., Adler, E., and Knowles, J.R. (1995). Phosphoglycerate kinase and triosephosphate isomerase from the hyperthermophilic bacterium *Thermotoga maritima* form a covalent bifunctional enzyme complex. EMBO J. *14*, 442–451.

Sharma, V., Prère, M.-F., Canal, I., Firth, A.E., Atkins, J.F., Baranov, P.V., and Fayet, O. (2014). Analysis of tetra- and hepta-nucleotides motifs promoting -1 ribosomal frameshifting in *Escherichia coli*. Nucleic Acids Res. *42*, 7210–7225.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., and Inouye, M. (1966). Frameshift mutations and the genetic code. Cold Spring Harb. Symp. Quant. Biol. *31*, 77–84.

Szamecz, B., Boross, G., Kalapis, D., Kovács, K., Fekete, G., Farkas, Z., Lázár, V., Hrtyan, M., Kemmeren, P., Koerkamp, M.J.A.G., et al. (2014). The Genomic Landscape of Compensatory Evolution. PLOS Biol *12*, e1001935.

Tamas, I., Wernegreen, J.J., Nystedt, B., Kauppinen, S.N., Darby, A.C., Gomez-Valero, L., Lundin, D., Poole, A.M., and Andersson, S.G.E. (2008). Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. Proc. Natl. Acad. Sci. *105*, 14934–14939.

Tu, A.H., and Turnbough, C.L. (1997). Regulation of *upp* expression in *Escherichia coli* by UTP-sensitive selection of transcriptional start sites coupled with UTP-dependent reiterative transcription. J. Bacteriol. *179*, 6665–6673.

Turnbough, C.L. (2011). Regulation of gene expression by reiterative transcription. Curr. Opin. Microbiol. *14*, 142–147.

Uptain, S.M., Kane, C.M., and Chamberlin, M.J. (1997). Basic mechanisms of transcript elongation and its regulation. Annu. Rev. Biochem. *66*, 117–172.

Viguera, E., Canceill, D., and Ehrlich, S.D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. *20*, 2587–2595.

Volchkov, V.E., Becker, S., Volchkova, V.A., Ternovoj, V.A., Kotov, A.N., Netesov, S.V., and Klenk, H.D. (1995). GP mRNA of Ebola virus is edited by the Ebola virus polymerase and by T7 and vaccinia virus polymerases. Virology *214*, 421–430.

Volchkov, V.E., Volchkova, V.A., Muhlberger, E., Kolesnikova, L.V., Weik, M., Dolnik, O., and Klenk, H.D. (2001). Recovery of infectious Ebola virus from complementary DNA: RNA editing of the GP gene and viral cytotoxicity. Science *291*, 1965–1969.

Wagner, L.A., Weiss, R.B., Driscoll, R., Dunn, D.S., and Gesteland, R.F. (1990). Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli.* Nucleic Acids Res. *18*, 3529–3535.

Yang, F., Moss, L.G., and Phillips, G.N. (1996). The molecular structure of green fluorescent protein. Nat. Biotechnol. *14*, 1246–1251.

Zhou, Y.N., Lubkowska, L., Hui, M., Court, C., Chen, S., Court, D.L., Strathern, J., Jin, D.J., and Kashlev, M. (2013). Isolation and characterization of RNA polymerase *rpoB* mutations that alter transcription slippage during elongation in *Escherichia coli*. J. Biol. Chem. *288*, 2700–2710.

CHAPTER 5
Discussion

Conclusion

RNA processing systems such as splicing and editing are known for their seemingly gratuitous complexity. Since its discovery, some have suggested that RNA editing may have evolved to modulate genetic sequences at the RNA level or increase evolutionary variation (Cavalier-Smith, 1997; Speijer, 2006). Alternatively, the theory of constructive neutral evolution (CNE) suggests that RNA editing emerged in a neutral fashion despite offering no intrinsic benefit (Covello and Gray, 1993; Stoltzfus, 1999). Although this idea of the non-adaptive emergence of RNA editing is plausible, it has yet to be experimentally tested and the main objective of this thesis was to assess the role of drift in the emergence of editing-type processes. The results presented in this body of work suggest a non-adaptive evolution of slippage-type editing processes in bacteria, and this is consistent with the neutral model for the evolution of RNA editing by Covello and Gray (1993).

Slippage-type editing is a process analogous to RNA editing where the sequence of the transcript differs from its encoding template. We have shown that RNA polymerase slippage results in the editing of information stored in the *E. coli* genome (Chapters 3 and 4). Sequencing of mRNA transcripts revealed that slippage corrected frameshift mutations observed at the genomic level, leading to expression of full-length proteins, but in turn reduced the efficiency of expression for in-frame genes bearing polyA/T sequences (Chapter 4). Notably, we only observed the emergence of 22 editing sites constituting approximately 0.5% of *E. coli* coding

genes, and, further to this, these sites were located in genes categorised as non-essential (Chapter 3). These results support the view that slippage-type editing may have evolved non-adaptively, and that slippage and editing do not compensate for the effect of Muller's ratchet. The result is a complex, yet inefficient, gene expression system.

Minor additive experiments

In Chapter 2, we analysed the impact of mutations accumulated during our bottleneck experiment on protein function using the delta-bitscore (DBS) method (Wheeler et al., 2016), but the results do not correspond directly to the loss of fitness observed. Although this method provided some insights into the severity of the mutational changes on protein function, we have not assessed non-coding and untranslated regions. Studies have shown that small RNAs (Gottesman, 2004; Waters and Storz, 2009) and untranslated cis-regulatory regions (Mandal and Breaker, 2004) can play major roles in the regulation of gene expression in bacteria. Therefore, analysing the effect of mutational changes in such regions could shed light on the observed loss of fitness. These could be assessed by recreating the observed mutations in otherwise unmutated (ancestor) strains. Additionally, to test the phenotypic impact of severe mutations, we could introduce the observed mutations back into the wild-type ancestor and subsequently compete the knock-in and wild-type lines to determine relative fitness (Barrick and Lenski, 2013). This will provide an indication as to whether the genes with high DBS contributed to the observed fitness loss. Further to this, we could perform whole transcriptome sequencing to assess the expression levels of genes, and using BioLog plates, we may be able to examine how and whether these genotypic changes affect phenotype.

The introduction of the observed frameshift mutation in *araC* from our mutation accumulation (MA) experiment showed that production of functional AraC is essential under conditions

where arabinose is provided as the sole carbon source (Chapter 3). Although we observed a reduction of growth rate in our knock-in, and subsequent mRNA sequencing showed transcript heterogeneity, we did not directly examine whether full-length AraC was being produced. In order to investigate whether full-length AraC is produced, we could perform western blotting using anti-AraC antibody and quantify the amount of AraC using mass spectrophotometry. While we have previously shown that cells bearing this frameshift *araC* gene were capable of metabolising arabinose, we have not investigated whether compensatory or back mutations could have corrected the frameshift at the genome level rather than through the action of slippage-type editing. We can assess this by sequencing the introduced frameshifted *araC* gene to check for reversions and perform whole-genome sequencing to investigate whether compensatory mutations have appeared. These will allow us to elucidate whether mutational and genotypic changes to the *araC* gene may have contributed to the metabolism of arabinose.

Previously, we measured and compared the growth rates of wild-type REL607 and REL607::*araC* 8T knock-ins in arabinose-only media, and showed that the knock-in had a lower growth rate compared to the wild-type (Chapter 3). The growth rates were calculated based on the minimum doubling time, however, and the maximum growth rate is only one component of fitness (Vasi et al., 1994). In order to better compare fitness between lines, we can quantify fitness as the ratio of the realised growth rates of two populations while they compete for resources in the same flask. A method that utilises fluorescent markers may be used to compete and distinguish between the strains in competition allowing us to measure fitness more accurately (Gullberg et al., 2011).

We have shown that tracts of As are prone to RNA polymerase slippage but studies on translational errors suggest that polyA/T tracts are also prone to ribosomal frameshifting

(Sharma et al., 2014). In order to investigate whether ribosomal frameshifting could have contributed to the reduction of full-length protein production, we could engineer a control GFP reporter system with an identical slippage-prone polyA tract but with the inclusion of a Shine-Dalgarno-like sequence upstream and a pseudoknot sequence downstream of the polyA tract. This will enable us to determine whether the reduction in protein production is a direct result of RNA polymerase slippage or whether it could have been an artefact of ribosomal frameshifting.

Furthermore, although our western blot results showed bands that corresponded to the predicted GFP protein, we have not provided concrete evidence that the bands are indeed the predicted GFP and we have not measured the protein concentration. We could perform shotgun discovery and selected reaction monitoring (SRM) mass spectrometry (Lu et al., 2009). This will enable us to determine the protein sequences produced and also accurately determine the relative amount of full-length and truncated GFP proteins produced. As a whole, these experiments would provide support for the works completed in this thesis.

Further discussion

The origins of RNA editing and complex molecular machines have been the topic of much debate, and are commonly believed to be the product of selection where an increase in complexity results in improved function (Bonner, 1988). Indeed RNA editing has been widely related to as a process that generates diversity (Landweber and Gilbert, 1993; Pullirsch and Jantsch, 2010) and although this may rationalize maintenance, it cannot easily explain its origin. Most fail to realise that RNA editing is essentially a mechanism of error correction, where RNA editing retailors a non-functional transcript producing a translatable transcript (Shaw et al., 1988; Simpson et al., 2000). RNA editing is particularly prevalent in organelles,

and it has been suggested that editing is a response to mutational pressures from the operation of Muller's ratchet in organellar genomes (Börner et al., 1997). We have shown that slippage-type editing effectively correct errors in genes which had acquired natural frameshift mutations, but in turn reduces expression efficiency of genes with intact reading frames. This is rather puzzling because the existence of RNA editing seems to offer no inherent advantage. Contrary to Börner et al. (1997), editing does not appear to be an effective way of escaping Muller's ratchet as RNA polymerase slippage yields dud mRNAs from polyA tracts before the emergence of frameshift mutations in those tracts. This type of ordering puts the problem before the solution: genomes accumulating deleterious mutations would be at a selective disadvantage in the absence of a mutation correction system. In other words, it is not plausible that mutations would be fixed in populations before the emergence of a corrective machine that corrects these mutations. As opposed to the adaptationist thinking, in editing it is not necessary for there to be a selective advantage for fixation of editing, it may simply arise through suitable preconditions. Stoltzfus (1999) notes that the recruitment of editing machinery may be explained by tinkering which involves pre-existing autonomous enzymes that are known in other functions (Jacob, 1977).

As shown in this thesis, the evolution of slippage-type editing is consistent with CNE (Stoltzfus, 1999), an extension to the model for the evolution of RNA editing (Covello and Gray, 1993). In our study, we showed that slippage-type editing emerged non-adaptively, before there was any need for editing. Upon fixation of frameshift mutations by genetic drift, slippage-type editing becomes indispensable for expression of functional genes. The capacity to reverse potentially deleterious mutations at the RNA level relaxes functional constraints at the gene level, allowing slightly deleterious mutations to accumulate. Thus, editing is only tolerated in the presence of the ratchet; it might be disadvantageous in the absence of the ratchet

(causes increased mutation). Such a series of events leads to the evolution of a complex system (Gray et al., 2010). However, complexity is not necessarily indicative of progress, as encapsulated in the theory of CNE. Complex machineries may have been acquired through CNE accretion, and not selective processes (Stoltzfus, 1999), where they can be subsequently co-opted for secondary functions under positive selection (Covello and Gray, 1993). The essence of CNE is that there does not have to be an adaptive explanation for all this extra complexity. It can be easily explained by neutral evolution in small populations, as seen in our mutation accumulation experiment, and this is particularly true for mitochondrial genomes which fit the prerequisite of Muller's ratchet having little to no genetic recombination and small population sizes (Kurland, 1992; Lynch, 1996, 1997). A brief discussion concerning the mitochondrial origins of splicing is helpful in order to understand the role of CNE in the origins of RNA editing.

The origin and subsequent evolution of the mitochondrion, powerhouse of the eukaryote cell has been long associated with eukaryotic cell evolution. Although much uncertainty surrounds the evolution of eukaryotes, most agree that the mitochondria evolved from an engulfed cell co-opted as an organelle (Gray et al., 1999; Poole and Gribaldo, 2014). Splicing is an example of a complex system that has been hypothesised to have originated in the mitochondria. The evolution of splicing could be explained by endosymbiont to host transfer of a group II intron followed by complexification to form the modern spliceosome. The mitochondrial seed hypothesis (Cavalier-Smith, 1991; Logsdon, 1998) stipulates that α-proteobacterial endosymbiont that evolved into the present-day mitochondria transported self-splicing group II introns into the eukaryotic host cell and were transferred via DNA-intermediates into the nuclear genome. They were subsequently spread as mobile genetic elements throughout nuclear chromosomes, thereafter the ribozyme structure degenerated and fragmented into

snRNAs that could facilitate splicing in *trans* (Sharp, 1991). Reversal of these fragmented introns would be incredibly rare, and in this way it is a ratchet-like process, where no positive selection is necessary. Sharp's theory of the division of a group II intron into "five easy pieces" fits the characteristics of a trait evolving by CNE.

Splicing and editing are similar in many ways; both are known to be complex cellular systems, they both modify pre-mRNA prior to the production of mRNA and both may generate multiple products from a single mRNA (Herbert and Rich, 1999). As aforementioned, splicing may have evolved as a result of endosymbiotic gene transfer (Cavalier-Smith, 1991; Logsdon, 1998), and the characteristics that splicing and RNA editing share suggest that it is plausible that RNA editing may also evolved in such a way. Mitochondria are naturally subjected to bottlenecks, much like the experimental conditions of this thesis, therefore it is plausible that editing may have emerged in the mitochondria and was subsequently transferred and fixed in the nucleus. Evidently, a myriad of RNA editing systems have been discovered in eukaryotic organelles (chloroplasts and particularly mitochondria) (Gray, 2003) compared to the very few documented in the nucleus. The origin of slippage-type editing in our study fits the mitochondrial origin of splicing, where traits that have evolved through non-selective forces in the mitochondrial genome could subsequently be transferred to the nuclear genome, where they obtain a secondary function. There is no reason to imply a selective advantage for the origin of splicing and RNA editing. These traits could instead evolve by CNE, and any subsequent function acquired by a trait could provide a selective advantage to an organism.

Prior to the evolution of sex, CNE could have operated in the nucleus (nuclear DNA compartment) in small populations. For example, in the absence of sex, neutral or slightly deleterious mutations can become fixed (at a rate equal to the mutation rate) in small

populations, leading to the origin of new traits that are not under selection. However, even in modern eukaryotic cells, where sex is believed to mitigate the effects of Muller's ratchet, it is possible for traits to evolve by CNE. Such an example might be seen in a population of early predators, assuming that the first eukaryotes were phagotrophs that consumed other cells.
In such a case, while sex might lessen the burden of slightly deleterious mutations, it cannot completely eliminate them, as seen in human populations (Graur, 2017). Thus, in any population where the effective population size is not high enough to counteract the effects of drift even in the presence of sex, one might expect mutations to accumulate, such that traits could evolve by CNE.

To conclude, even though sexual populations undergo recombination which aids in the removal of deleterious mutations, the reality is that often population sizes are not large enough for many mutations to be eliminated from a population. If this is taken into consideration, we could theoretically evolve RNA editing in sexual eukaryotes with a small population size, like that observed in a small asexual population in our mutation accumulation experiment. Even though sex may eliminate a proportion of deleterious mutations and prevent an increase in mutational load, under conditions of strong drift, we would expect to see the accumulation of mutations, as recombination cannot entirely eliminate the effect of Muller's ratchet. Under these conditions, RNA editing may evolve as a system that allows a wider array of mutations to be sublethal, so that mutations that would normally be lethal without editing are now tolerated. Few studies have experimentally tested the evolution of RNA editing in eukaryotes and the bigger question of the origins of eukaryotic complexity remains unanswered. The combination of experimental evolution and genome sequencing will enable us to test the evolutionary origins of RNA editing which may elucidate a greater understanding of how complex traits of eukaryotic cells may have evolved as a result of genetic drift.

References

Barrick, J.E., and Lenski, R.E. (2013). Genome dynamics during experimental evolution. Nat.

Rev. Genet. *14*, 827–839.

Bonner, J.T. (1988). The Evolution of Complexity by Means of Natural Selection (Princeton, N.J: Princeton University Press).

Börner, G.V., Yokobori, S., Mörl, M., Dörner, M., and Pääbo, S. (1997). RNA editing in metazoan mitochondria: staying fit without sex. FEBS Lett. *409*, 320–324.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. Trends Genet. *7*, 145–148.

Cavalier-Smith, T. (1997). Cell and genome coevolution: facultative anaerobiosis, glycosomes and kinetoplastan RNA editing. Trends Genet. TIG *13*, 6–9.

Covello, P.S., and Gray, M.W. (1993). On the evolution of RNA editing. Trends Genet. TIG *9*, 265–268.

Gottesman, S. (2004). The small RNA regulators of Escherichia coli: roles and mechanisms*. Annu. Rev. Microbiol. *58*, 303–328.

Graur, D. (2017). An Upper Limit on the Functional Fraction of the Human Genome. Genome Biol. Evol. *9*, 1880–1885.

Gray, M. (2003). Diversity and Evolution of Mitochondrial RNA Editing Systems. IUBMB Life *55*, 227–233.

Gray, M.W., Burger, G., and Lang, B.F. (1999). Mitochondrial evolution. Science *283*, 1476–1481.

Gray, M.W., Lukeš, J., Archibald, J.M., Keeling, P.J., and Doolittle, W.F. (2010). Irremediable Complexity? Science *330*, 920–921.

Gullberg, E., Cao, S., Berg, O.G., Ilbäck, C., Sandegren, L., Hughes, D., and Andersson, D.I. (2011). Selection of Resistant Bacteria at Very Low Antibiotic Concentrations. PLOS Pathog. *7*, e1002158.

Herbert, A., and Rich, A. (1999). RNA processing and the evolution of eukaryotes. Nat. Genet. *21*, 265–269.

Jacob, F. (1977). Evolution and tinkering. Science *196*, 1161–1166.

Kurland, C.G. (1992). Evolution of mitochondrial genomes and the genetic code. BioEssays News Rev. Mol. Cell. Dev. Biol. *14*, 709–714.

Landweber, L.F., and Gilbert, W. (1993). RNA editing as a source of genetic variation. Nature *363*, 179–182.

Logsdon, J.M. (1998). The recent origins of spliceosomal introns revisited. Curr. Opin. Genet. Dev. *8*, 637–648.

Lu, B., Xu, T., Park, S.K., and Yates, J.R. (2009). Shotgun protein identification and quantification by mass spectrometry. Methods Mol. Biol. Clifton NJ *564*, 261–288.

Lynch, M. (1996). Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. Mol. Biol. Evol. *13*, 209–220.

Lynch, M. (1997). Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. Mol. Biol. Evol. *14*, 914–925.

Mandal, M., and Breaker, R.R. (2004). Gene regulation by riboswitches. Nat. Rev. Mol. Cell Biol. *5*, 451–463.

Poole, A.M., and Gribaldo, S. (2014). Eukaryotic Origins: How and When Was the Mitochondrion Acquired? Cold Spring Harb. Perspect. Biol. *6*, a015990.

Pullirsch, D., and Jantsch, M.F. (2010). Proteome diversification by adenosine to inosine RNA editing. RNA Biol. *7*, 205–212.

Sharma, V., Prère, M.-F., Canal, I., Firth, A.E., Atkins, J.F., Baranov, P.V., and Fayet, O. (2014). Analysis of tetra- and hepta-nucleotides motifs promoting -1 ribosomal frameshifting in Escherichia coli. Nucleic Acids Res. *42*, 7210–7225.

Sharp, P.A. (1985). On the origin of RNA splicing and introns. Cell *42*, 397–400.

Sharp, P.A. (1991). Five easy pieces. Science *254*, 663.

Shaw, J.M., Feagin, J.E., Stuart, K., and Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell *53*, 401–411.

Simpson, L., Thiemann, O.H., Savill, N.J., Alfonzo, J.D., and Maslov, D.A. (2000). Evolution of RNA editing in trypanosome mitochondria. Proc. Natl. Acad. Sci. *97*, 6986–6993.

Speijer, D. (2006). Is kinetoplastid pan-editing the result of an evolutionary balancing act? IUBMB Life *58*, 91–96.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. J. Mol. Evol. *49*, 169–181.

Vasi, F., Travisano, M., and Lenski, R.E. (1994). Long-Term Experimental Evolution in Escherichia coli. II. Changes in Life-History Traits During Adaptation to a Seasonal Environment. Am. Nat. *144*, 432–456.

Waters, L.S., and Storz, G. (2009). Regulatory RNAs in Bacteria. Cell *136*, 615–628.

Wheeler, N.E., Barquist, L., Kingsley, R.A., and Gardner, P.P. (2016). A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. Bioinforma. Oxf. Engl.